Assignment 1

Shreya Sakura Noskor - 1007996563
Ben Marlow - 1006724920
Timothy Pasaribu - 1009714864
Huayin Luo - 1007864517

Part 1 - Interest statements:

In the 21st century, we have databases full of scientific journals and papers that have brought new insights into various topics. From medical research papers highlighting new drugs to psychology papers emphasizing the importance of human behaviour, most students and professionals are overwhelmed by the amount of knowledge being released every year. This makes it nearly impossible for them to stay current and results in a "lag" where they may fall behind on new findings in their field.

Our team, being part of the student demographic, understands the very real struggle of reading 10 different 30-page research papers and finding that some information overlaps while others contradict each other. We want to build a tool that doesn't just summarize research, but contextualizes it within the larger landscape of the topic users are studying.

Individually, there are parts of the project we are each excited to work on:

Sakura: Being a student, we have to read a lot of papers for classes, and some of the papers either start overlapping or contradicting one another. I'm looking forward to working on the summarization and parsing side of the platform, helping extract key findings from papers and clearly showing where results overlap or contradict across studies.

Ben: Studying both physics and computer science, it can be difficult to properly read and understand papers from different fields, which often have completely different structures and scopes. I am very excited to build a tool that could help me make sense of the similarities and differences and how the research in these fields overlaps.

Timothy: As a data science student and ML research assistant, I often read papers across ML and NLP. Even when I understand them individually, it's hard to see how they connect or fit into the bigger picture. I'm excited to contribute to building the relationship-mapping and contextualization aspects of the tool, connecting papers to show how the broader research landscape.

Huayin: It's often super hard to navigate the quickly changing fields of research, and I've learnt as a student and heard from graduate students/researchers that one of the hardest but also most important things is to keep up to date with the current SOTA. I'm looking forward to working on features that help users stay up to date with state-of-the-art research, especially within fast-moving niches, by organizing and surfacing the most relevant and recent developments.

Part 2 - Landscape Analysis:

Table 1: Eight relevant companies, two open source projects and one research paper

Relevant Item	Description	Commentary
OpenEvidence https://www.openevidence.com/	A specialized AI platform for medical professionals that provides cited answers from 30M+ PubMed papers.	Highly accurate for Q&A, but lacks the visual "live map" of a field that our project aims to build for broader discovery.
Consensus https://consensus.app/	An AI academic search engine that synthesizes "yes/no" answers from research papers and provides "SciScore" quality metrics.	good for binary analysis (e.g., "does this drug work? yes/no"), but doesn't show the structural relationships between side effects or secondary findings.
Elicit https://elicit.com/	An AI research assistant focusing on automating systematic reviews and data extraction into custom tables.	A linear workflow tool; our project adds a non-linear graph view to see connections between different reviews.
Connected Papers https://www.connectedpapers.com/	A visual discovery tool that uses citation proximity to build 2D similarity graphs of academic papers.	Purely a discovery tool. Doesn't "read" the papers for the user or extract clinical insights like side effects
ResearchRabbit https://www.researchrabbit.ai/	A "discovery engine" that allows users to create collections and maps citation networks over time.	This has the map part, but summaries are basic abstracts. Also, the map is purely based on citations, but not all papers cite each other directly. We could add some sort of semantic component?
PrimeKG (Open Source) https://github.com/mims-harvard/PrimeKG	A precision medicine knowledge graph that integrates many resources covering diseases, drugs, and pathways.	We can use PrimeKG as a ground-truth "backbone" to verify the semantic connections our AI makes between clinical papers.
LangGraph (Open Source) https://www.langchain.com/langgraph	A library for building stateful, multi-agent AI workflows, specifically for creating complex RAG loops.	We could use this to scale our project up, since there's so many papers out there and sources of information. LangGraph is the industry standard way to implement GraphRAG, and could help us maintain a

		persistent state.
Scite AI https://scite.ai/	A citation analysis tool that classifies references as "supporting," "mentioning," or "contradicting" the original claim	The part about differing views is pretty cool, maybe we could incorporate this into our knowledge graph.
Paperguide https://paperguide.ai/	An all-in-one AI assistant combining literature reviews, reference management, and "Chat with PDF" features.	It doesn't have the knowledge graph feature, but its scope / framing as an "all in one research-assistant" could overlap with our target audience
Cypris AI https://www.cypris.ai/	An enterprise-level R&D intelligence platform that combines scientific literature with patent data analysis.	More for enterprise-level rather than the individual user. Can see what the difference is between these markets. Focus on *real-time* innovation data and actionable insights, which is part of our strategy too.
Multi-Document Scientific Summarization from a Knowledge Graph-Centric View https://arxiv.org/abs/2209.04319	Introduces KGSum, a Multi-Document Scientific Summarization (MDSS) which first generates a knowledge graph by extracting entities and relations from the paper. It uses abstracts of the related papers, and full text of source paper.	This talks about how papers need to be connected by topic instead of the citations within them to show " coherent and concise summaries" for various topics. It uses a two-stage decoder so that it first generates a summary graph, before the actual summary. This could be a robust way to prevent hallucinations.

Part 3 - Project Outline:

Problem statement

The volume of new papers makes it hard for researchers to maintain an up-to-date, coherent picture of a field, especially in fast-moving areas like AI and Medical Science. Existing tools either focus on summarizing single PDFs or on high-level literature mapping, but rarely integrate deep paper understanding with personalized next-paper recommendations tied to concrete research goals. As a result, researchers waste time skimming marginally relevant papers, duplicate prior work, and struggle to build well-structured related-work sections and project directions efficiently.

Proposed solution

We build an AI-assisted reading environment that ingests a paper (PDF/URL/in-web-selection) and generates structured artifacts such as core claims and methods, limitations, and takeaways. On top of this, we construct an interactive knowledge graph that links the current paper to prior and subsequent work (using citations and semantic similarity), and recommends what to read next based on recency,

topical relevance, and user-specified research objectives. The system aims to help users both read a given paper faster and continuously curate a high-value reading path.

High-level technical approach and milestones

Milestone 1 – Core ingestion and parsing : Implement robust PDF ingestion (figure/latex stripping, section detection) and possibly a metadata retrieval via APIs such as Semantic Scholar (title, abstract, references, citations).

Milestone 2 – Paper understanding and summarization: Use LLM-based pipelines and prompt templates to extract research questions, main hypotheses/claims, methods, key results, and limitations. Evaluate quality with small human-annotated sets (e.g., does the tool capture the desired "1–3 key points/claims").

Milestone 3 – Knowledge graph and recommendation engine: Build a graph where nodes are papers and edges encode citations, co-citation, bibliographic coupling, and semantic similarity using embeddings from titles, abstracts, and other representative sections. Implement ranking that combines graph-based relevance, recency, and user constraints for next-paper suggestions.

Unknowns to investigate

- Summary reliability: Ensuring extracted claims and results remain grounded in the original paper and avoid hallucinations, especially for technical details.
- Evaluation metrics: Measuring whether the system improves literature review quality or time-to-insight compared to baselines such as keyword search or citation graphs.
- Data coverage: Assessing limitations of open citation APIs such as *Semantic Scholar* (e.g., coverage, rate limits, etc.) for large-scale graph construction.

Part 4 - Press Release:

Scientists, students, and medical professionals today face a daunting challenge: making sense of a rapidly expanding body of literature which is simply impossible to keep up with. Thousands of new papers are published every week, making it increasingly difficult for researchers to stay up-to-date on the latest and most relevant information being published in their fields. Even worse: as research becomes more interdisciplinary, scientists and clinicians are expected to understand a wider range of topics than ever before, making it easy to overlook major gaps in their own knowledge.

Today, we announce MindMap, an interactive application that uses machine learning to help users understand, summarize, and structurally reason about scientific and medical papers. Beyond aiding comprehension of specific papers, our application builds an interactive "MindMap" of a user's chosen research field, highlighting their areas of greatest awareness and more importantly: where the gaps may lie. MindMap will actively recognize where a user's knowledge is lacking, and suggest papers, textbooks, or online resources to gain that knowledge.

"Reading across different disciplines felt like learning five languages at once- and none of them fluently," said Natalie De Witt, an early user. *"MindMap shows me where I'm conversational and where I'm completely lost."*

Using advanced natural language processing, MindMap ingests PDFs of medical and scientific papers, and extracts methods, results, and key claims into a condensed summary which can be understood quickly and conveniently. MindMap allows users to tailor the depth and complexity of outputs, making cutting-edge research accessible to new students and training clinicians, even without prior background knowledge. The map itself is built interactively over time, allowing users to explore fields visually, dive deep into specific topics, and see which conclusions are strongly supported and which are still being debated.

For clinicians, this means faster insight into evolving medical evidence. For scientists, it means identifying under-researched fields and avoiding redundant work. For students, it provides a clear mental model of expansive topics that may be intimidating to conceptualize.

"As an emergency room physician, I don't have time to read every paper front-to-back," said Dr. Daniel Ngui. *"The mind map shows me where consensus exists and where uncertainty remains. That's incredibly valuable when making decisions for treatment in unusual cases."*

MindMap is currently in early access, with plans to expand scope and support very soon.

Appendix

Press release iteration 1: more technical, targeted at an audience with machine learning experience

Scientists, students, and medical professionals today face a daunting challenge: making sense of a rapidly expanding body of literature which is simply impossible to keep up with. Thousands of new papers are published every week, making it increasingly difficult for researchers to stay up-to-date on the latest and most relevant information being published in their fields. Even worse: as research becomes more interdisciplinary, scientists and clinicians are expected to understand a wider range of topics than ever before, making it easy to overlook major gaps in their own knowledge.

Today, we announce MindMap, an interactive platform utilizing advanced Natural Language Processing to help users summarize and comprehend scientific and medical literature. MindMap will actively recognize where a user's knowledge is lacking, and suggest papers, textbooks, or online resources to gain that knowledge.

"Reading across different disciplines felt like learning five languages at once— and none of them fluently," said Natalie De Witt, an early user. *"MindMap shows me where I'm conversational and where I'm completely lost."*

MindMap makes use of advanced Transformer-based architectures to ingest scientific PDFs and perform named entity recognition and relation extraction. It condenses methodology, results, and primary claims into compressed summaries. MindMap allows users to tailor the depth and complexity of outputs, making cutting-edge research accessible to new students and training clinicians, even without prior background knowledge. The map itself is built interactively over time, allowing users to explore fields visually, dive deep into specific topics, and see which conclusions are strongly supported and which are still being debated.

For clinicians, MindMap offers a low-latency path to evidence-based insights. For researchers, it serves as a tool for novelty detection, identifying under-explored nodes in the research graph to avoid redundant work. For students, it provides a structured mental model, accelerating the process of starting fresh in a new technical field.

“In a clinical environment, the signal-to-noise ratio in new literature is often too low for real-time application,” said Dr. Daniel Ngui. “The MindMap interface clarifies where the evidence converges and where the uncertainty remains, allowing for better-informed Bayesian reasoning in patient care.”

MindMap is currently in early access, with plans to expand scope and support very soon.

Press release iteration 2: less technical, intended for audiences with no knowledge of ML terminology

Scientists, students, and medical professionals today face a very difficult challenge: making sense of a massive amount of papers that are impossible to keep up with. Thousands of new papers are published every week, making it increasingly hard for researchers to stay up-to-date on the latest and most relevant information being published in their fields. Even worse: as research becomes more broad and encompasses different topics, scientists and clinicians are expected to understand a wider range of topics than ever before, making it easy to overlook major gaps in their own knowledge.

Today, we announce MindMap, an interactive application that uses machine learning to help users understand, summarize, and structurally reason about scientific and medical papers. Beyond aiding learning, our application builds an interactive “MindMap” of a user’s chosen research field, highlighting their areas of greatest awareness and more importantly: where the gaps may lie. MindMap will actively recognize where a user’s knowledge is lacking, and suggest papers, textbooks, or online resources to gain that knowledge.

“Reading across different disciplines felt like learning five languages at once— and none of them fluently,” said Natalie De Witt, an early user. *“MindMap shows me where I’m conversational and where I’m completely lost.”*

Using state-of-the-art technology, MindMap works by taking document files of medical and scientific papers, and extracts methods, results, and key claims into a shortened summary which can be understood quickly and conveniently. MindMap allows users to change the depth and complexity of outputs, making advanced research accessible to new students and training physicians, even without prior background knowledge. The map itself is built interactively over time, allowing users to explore fields themselves, dive deep into specific topics, and see which conclusions are strongly supported and which are still being debated.

For clinicians, this means faster insight into evolving medical evidence. For scientists, it means identifying under-researched fields and avoiding redundant work. For students, it provides a clear mental model of expansive topics that may be intimidating to conceptualize.

"As an emergency room physician, I don't have time to read every paper front-to-back," said Dr. Daniel Ngui. "*The mind map shows me where consensus exists and where uncertainty remains. That's incredibly valuable when making decisions for treatment in unusual cases.*"

MindMap is currently in early access, with plans to expand scope and support very soon.

Iteration 3 is the final version used in Part 4. It blends elements of the first and second iterations to find a balance between accessible and technical writing.