1) Problem & Goal

Problem: The volume of AI/LLM papers has exploded; topic lists by year don't reveal influence, key actors, or direction of change.
Goal: Build a directed paper-citation network (2015–present) from top universities (QS top-10 in AI/Data Science) to answer: who is most influential, which subtopics are densest, which institutions/countries lead output, and where new work is heading.

2) Research Questions:

1. Which paper is the most impactful?
2. Oldest paper that's still very relevant until now?
3. What subtopics have the highest concentration of research?
4. Which country/institution has the most research output?
5. Where is the newer research headed?

3) How to collect data:

Primary sources: OpenAlex / Semantic Scholar (include citations and affiliations).
Years: 2015–present.
Topics: AI/NLP/LLM/Foundation Models/Transformers (keyword OR + subject filters).
Institutions: QS top-10 in AI/DS (filter by name or ROR/GRID).
Types: journals + top conferences (ACL, NeurIPS, ICML, ICLR, EMNLP, AAAI, etc.).
Fields kept: id, title, authors, affiliations, country, year, venue, keywords, reference list and citation counts.
Cleaning: Normalize institution aliases; handle missing affiliations as "Unknown"; for multi-affiliation papers report both first-author and author-share results (sensitivity check).
Scale control: Target 2k–8k nodes; if larger, use k-core/threshold sampling or whitelist venues.

4) Definition of nodes and edges:

Node: Papers with subtopics will be differed by node color, and influence/impact of each paper will be differed by node size using the score of pagerank, paper ages with opacity.
Edges: citations between papers.

- For nodes A and B, if paper of node A has cited paper of node B at least once, then there will be a directed edge from A to B.
- In the preview case, A be considered as the authority node, and B is the hub node.

5) Metrics will be used:

Degree count: show authority and hub score of each node. (eg: which cite most, which be cited most, total citations number)
Betweenness Centrality: Show the location of the paper in the network. High betweenness of paper might be a potentially unifying idea. New papers with high betweenness centrality may answer where the newer research headed.
Closeness centrality: Shows how foundational is that paper, how many papers were written based on that particular paper. Old paper with high closeness centrality means that it is still relevant until now.

PageRank/Eigenvector Centrality: Show the influence of each paper (highly cited by other highly cited papers).
Density: Show the tightness of the network, and within its sub-networks.
Modularity: How partitioned the network is.

6) Analysis Plan

We will collect AI/LLM papers from 2015, clean and standardize them into a single dataset, and build a directed citation network in R. We'll summarize the network's basic structure and focus on the main component. Core centralities will highlight influential papers, and we'll report concise Top-N results. We will detect topic communities, label them, and compare how concentrated they are. We'll also aggregate by institution and country to show output and impact.

7) Operational Definitions

- "Most impactful" = primarily PageRank, supported by in-degree; also show external-citation share and recent-citation share.
- "Old but relevant" = early year + top-5% recent influence (PageRank/in-degree), and high closeness centrality.
- "Highest concentration subtopics" = communities with high internal density, strong modularity contribution, and large node share.
- Affiliation rule = default first-author; also report author-share variant.

8) Timeline (4 weeks)

W1 – Data: finalize keywords & institution list; fetch & clean, W2 – Graph & Baselines: stats, centralities, Top-N; first figures, W3 – Communities & Time: labels, trends, aggregation; draft results, W4 – Polish & Repro: error checks, finalize Rmd & slides.

9) Team & Roles

Members:

1. Jinxi Hu 48528608
2. Johanes Panjaitan 39809579
3. Yuxuan Sun 27929934
4. Samarth Grover 38220463

Even split & rotation. Tasks in each phase are divided into equal-sized subtasks; ownership and reviewer roles rotate weekly (facilitator/scribe/repo maintainer all rotate; no permanent leads). Collected data and analysis codes will be stored in a team github repository.