The Network of Modern AI Research: A Quantitative Analysis of Influential Publications and Research Communities (2023-2025)

Hesam Sheikh Hassani, Artificial Intelligence, 0001111590
Safoura Banihashemi, Artificial Intelligence, 0001109509
Shafagh Rastegari, Artificial Intelligence, 0001118088

Project Repository: https://github.com/hesamsheikh/daily-papers-analysis

1 Introduction

With the rise of large language models (LLMs) and transformer-based architectures, AI research has witnessed substantial growth. New influential papers are published daily by the research community, helping us understand more about these high-scale models. This dynamic research activity asks for a systematic analysis of the current state of AI advancements.
In this study, we explore the key research papers featured on the Hugging Face daily papers[1] page. We scaped the paper metadata, including paper details and authors' information, and utilized an LLM to extract the organizations behind these publications (e.g., universities and companies). Using this data, we constructed a graph dataset of this research network. Finally, we examined this network to gain valuable insights.

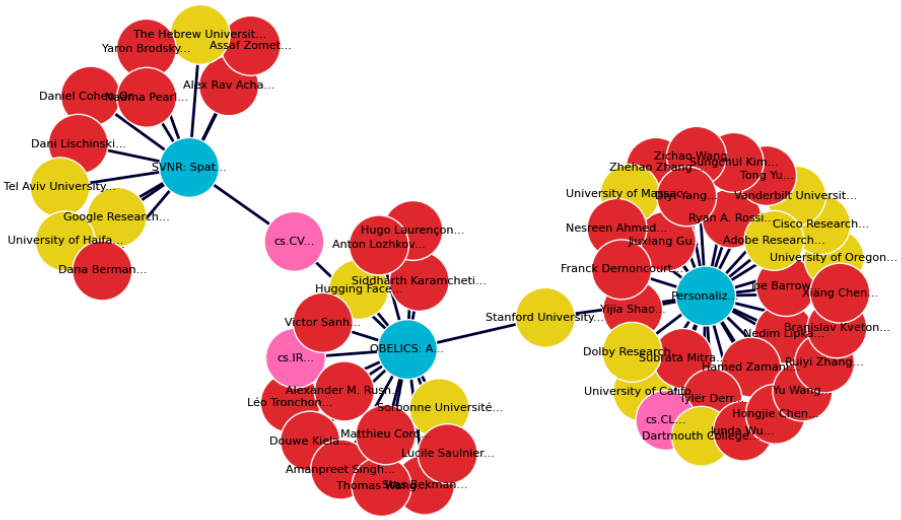

Figure 1: An example of a graph consisting of three paper nodes (cyan), their authors (red), organizations (yellow), and categories (pink).

2 Problem and Motivation

The exponential growth in AI research ignited by the advancements of LLMs has created a vast landscape. Because of the availability of LLMs and their novelty, the research in this field is widespread; numerous entities contribute to the research by building new models or exploring the existing ones. Additionally, the high computational costs and substantial funding required to train large-scale models have driven increased industry participation, leading to a level of collaboration between academia and industry rarely seen in other scientific fields [2].

This has resulted in a diverse body of work, making it challenging to track key contributors to this research landscape. As a result, understanding the structure of this landscape requires a systematic analysis to identify patterns of collaboration, institutional impact, and the flow of knowledge across various entities. The sheer volume of publications makes it difficult to identify which works are truly influential, which institutions drive major research, and how this research effort is distributed globally. Such a study can help different actors (e.g. countries) to identify how much they truly have an impact in leading or keeping up with these breakthroughs.

To address these challenges, this study aims to construct a structured network of AI research by analyzing papers featured on the Hugging Face Daily Papers page. Unlike studies that examine hundreds of thousands of publications, this research focuses on a selection of the most influential and high-impact papers that inspire a large community. By extracting metadata, institutional affiliations, and research topics, we build a graph dataset that provides a clearer picture of AI research dynamics. This helps us gain a deeper understanding of the field's evolution, helping researchers, industry professionals, and policymakers make more informed decisions about future directions, collaborations, and resource allocation.

3 Datasets

The Daily Papers page of Hugging Face lists the most important papers based on their impact. These papers include a variety of different topics in machine learning, computer vision, novel benchmarks, but mostly large language models. While this collection doesn't definitively cover all of the papers, it does feature the ones that are of high importance in the community. The collection does not include most of the metadata we need to create our dataset, necessitating the need to create our graph dataset and gather the key information we need.
We create this network of AI research in the following steps:

3.1 Listing Papers

We scrape this collection of papers in the time range of 10th May 2023 and 10 Feb 2025. This gives us a list of **7791** paper titles in that time frame of **643** days that are featured on the Daily Papers page. This process is not time-consuming, but rather straightforward.

3.2 Extracting Paper Metadata

With the list of top papers available, our next objective was to extract the metadata of the papers. We utilized the arxiv Python library to access the arXiv API [3]. Then we used this API to search for papers given their titles. ArXiv would then return the paper metadata which includes information such as the date of publishing, title, authors, summary, categories, the abstract, and

the PDF link of the paper. This search process took 2 hours and 25 minutes in total and would return **4195** paper objects, as the duplicate papers were removed and some titles did not yield results from the arxiv API. The collection of these papers is saved in a **32MB** JSON file that is accessible in the project repository.

We extract the authors and the subject category of each paper in this step from the metadata.

3.3 Extracting Publishing Organizations

The publishing organization of a paper is the universities, research centers, companies, or similar entities which the authors or the paper are affiliated with. As the arxiv API did not include any details as to which organizations contributed to the research paper, extracting this turned out to be challenging. We particularly analyzed two methods to extract this information:

1. Using the scholarly [4] Python library to search author names on Google Scholar[5] and find the organization they are affiliated with. However, due to the anti-crawling policies of Google Scholar, the search is intentionally time-consuming and thus impractical.

2. Searching for author information through ORCID public API [6]. This turned out ineffective as for many author names, multiple individuals with the same name would be returned.

Given these limitations, we took extra measures to extract the organizations' information without accessing public APIs, but by extracting them using LLMs. This is time-consuming and computationally expensive, but is the most reliable method for two reasons: the publicly available APIs may not return the most recent organization affiliations of a person, but the organizations mentioned in the paper are most definitely affiliated with the paper. Additionally, an individual may have multiple organization affiliations, and resolving which is affiliated with a given paper cannot be easily resolved.
Thus, we extract organization information from papers in two steps:

3.3.1 Downloading Papers

To parse the papers with LLM we used the PDF link from the paper metadata to download the full papers, resulting in 30GB of saved files. This process took several hours to complete.

3.3.2 LLM Extraction

We used Ollama[7] to serve LLMs locally. Ollama provides tools to download, run, and host various large language models with minimum effort. A local device with a Nvidia Geforce RTX 2060 was used to serve the LLM. For the limited computational budget and relative simplicity of the task at hand, we tested various lightweight LLMs and decided to process the PDFs using Qwen2.5-3B[8].

To limit the number of input tokens, we only extracted the text of each paper that would appear before the *Abstract* section. Processing all papers with the LLM took approximately 4 hours to complete, and the model in most cases was able to return the correct organization names, even though it sometimes detected false positives (such as the address of an organization).

Prompt

You will be provided with a segment of text extracted from the beginning of a research paper (before the abstract). Your task is to extract a list of unique organization names mentioned in the text. Follow these instructions exactly:

1. **Extract Only Organizations:** Identify and extract only the names of organizations. Do not include:
    - Addresses (e.g., lines that contain location details like "Bethesda, MD, USA")
    - Email addresses
    - Author names or other non-organization text
2. **Hints for Organization Names:** Organization names usually include keywords such as University, Institute, College, Laboratory, Center, Corporation, Inc., Ltd., etc.
3. **Formatting:**
    - List each organization on its own line.
    - Do not include any additional text, commentary, or punctuation.
    - If an organization name spans multiple lines, combine them into one single name (without the extra line breaks).
4. **Avoid Duplicates:** If an organization is mentioned more than once, list it only once.
5. **No Organizations Found:** If no organization names are identified, output exactly:

```
No organizations found
```

Now, process the following text:

3.4 Graph Dataset Format

We use the *CSV* format to save the graph datasets. Different entities and relationships between them are saved in separate files: paper, author, category, organization, paper-author relationship, paper-category relationship, and paper-organization relationship. This means we can create three graphs (using the relationships) and one whole graph that includes all entities and their relations. We can also infer additional graphs using these files (e.g. from the paper-author graph we can create an author-author graph in which co-authors are connected).

Type	Entity	Number
Node	Paper	4194
Node	Author	18523
Node	Category	224
Node	Organization	3669
Edge	Paper-Author	28675
Edge	Paper-Category	8601
Edge	Paper-Organization	9783
Edge	Author-Author	241405
Edge	Organization-Organization	14572

Table 1: The number of nodes and edges in the constructed graph datasets.

4 Validity and Reliability

The validity of the dataset used in this study is ensured by sourcing it from the Daily Papers page, a widely recognized platform that highlights high-impact AI research papers. This dataset prioritizes influential publications based on their reception within the research community. However, since it does not comprehensively cover all AI-related publications, selection bias may be present. While this dataset provides a faithful representation of AI research collaborations, certain limitations remain, such as missing metadata and evolving author affiliations, which may affect completeness.

The reliability of this study is addressed through a systematic and reproducible methodology for processing and analyzing AI research collaboration networks. The dataset was processed using pandas for data handling and networkx for network construction. The graph construction process followed a structured pipeline, such as duplicate removal. Network analyses, including degree centrality, modularity, and community detection, were conducted using deterministic algorithms where feasible. For stochastic methods, random seeds were explicitly set to ensure repeatability. While minor variations may arise due to API updates, the overall methodology guarantees a consistent and reliable representation of AI research networks.

Due to the probabilistic nature of LLMs, the extraction of organizations from the papers may yield slightly different results when running the process again or using a different large language model. Leveraging bigger models, potentially with reasoning capabilities ensures a more accurate result, but also higher computational demands. However, our qualitative survey of the extracted organizations shows that the overall process is reliable and trustworthy.

5 Measures and Results

By constructing and analyzing three distinct graphs The Core Network, Co-authorship Network, and Co-organization Network, we aim to uncover the intricate relationships and patterns that define their collaboration.

The Core Network, integrates authors, organizations, and papers, providing a holistic view

of the research ecosystem and highlighting key players and interactions. The **Co-authorship Network** focuses on the connections between authors and their publications, revealing patterns of co-authorship and individual contributions. Finally, the **Co-organization Network** examines the ties between organizations and papers.

Overall, these networks offer a multi-dimensional perspective on the research community, enabling us to identify influential authors, collaborative organizations, and quantitative analysis of the published papers. Through this analysis, we seek to better understand how knowledge is created, shared, and propagated within this research network.

5.1 The Core Network

In this section, we examine the Core Network, a structured graph that connects papers, authors, and organizations. Figure 2 presents a visualization of this network, showcasing the relationships between these entities and highlighting the most influential nodes.

We analyze the network's structural properties by analyzing its degree distribution, robustness, and key centrality measures. A highly skewed degree distribution suggests the presence of influential figures, while robustness analysis evaluates the network's resilience to node removal. Additionally, centrality measures such as degree centrality, betweenness centrality, and PageRank help identify key contributors within the AI research landscape.

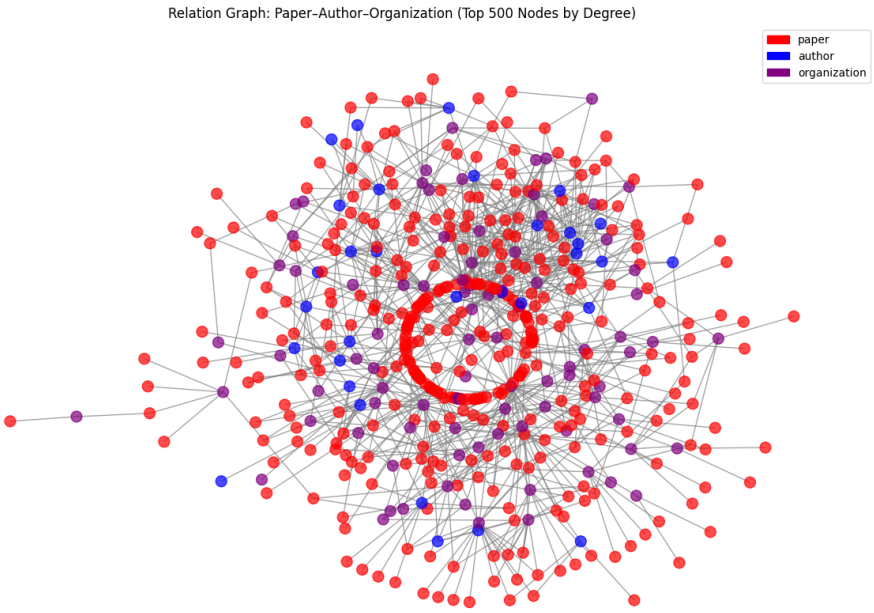

Figure 2: Paper-Author-Organization Collaboration Network (Top 500 Nodes by Degree)

5.1.1 Degree Distribution Analysis

The degree distribution of the network provides key insights into its structural properties and connectivity patterns. As shown in Figure 3, the distribution exhibits a highly skewed nature, where the majority of nodes have a low degree, while a small fraction maintains significantly higher connectivity. This pattern indicates a power-law distribution, a common characteristic

of scale-free networks. In such networks, a few highly connected hub nodes serve as central points for information flow, playing a crucial role in linking different parts of the network.

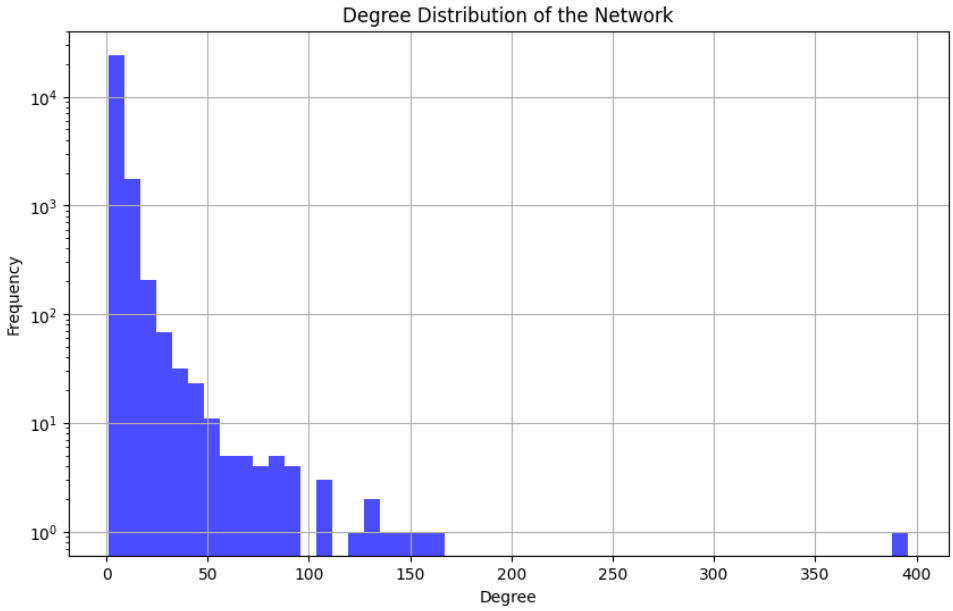

Figure 3: Degree Distribution Showing Scale-Free Nature of the Core Network

5.1.2 Scale-Free Network Characteristics

The degree distribution plot on a log-log scale, shown in Figure 4, further confirms the scale-free nature of the network. The near-linear trend suggests that node connectivity follows a power-law decay, where most nodes have relatively few connections, while a small number of highly connected nodes act as central hubs. This structure indicates that a set of papers, authors, and institutions disproportionately shape the research landscape.

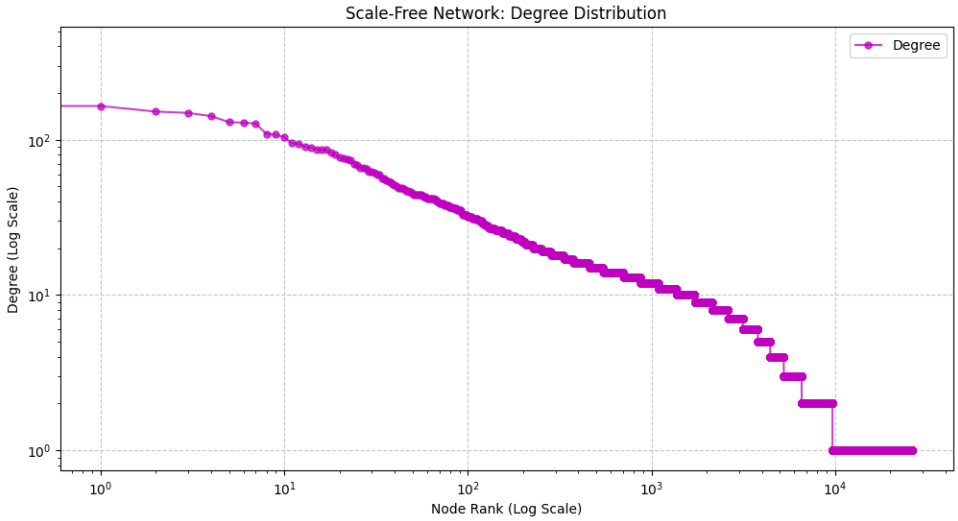

Figure 4: Log-Log Degree Distribution of the Research Network

5.1.3 Network Robustness Analysis

To evaluate the network's resilience, we conducted a robustness analysis by systematically removing nodes and measuring their impact on connectivity. As shown in Figure 5 the plot demonstrates a mostly linear decline in connectivity, indicating that the network maintains its overall structural integrity even after the removal of a significant portion of nodes.

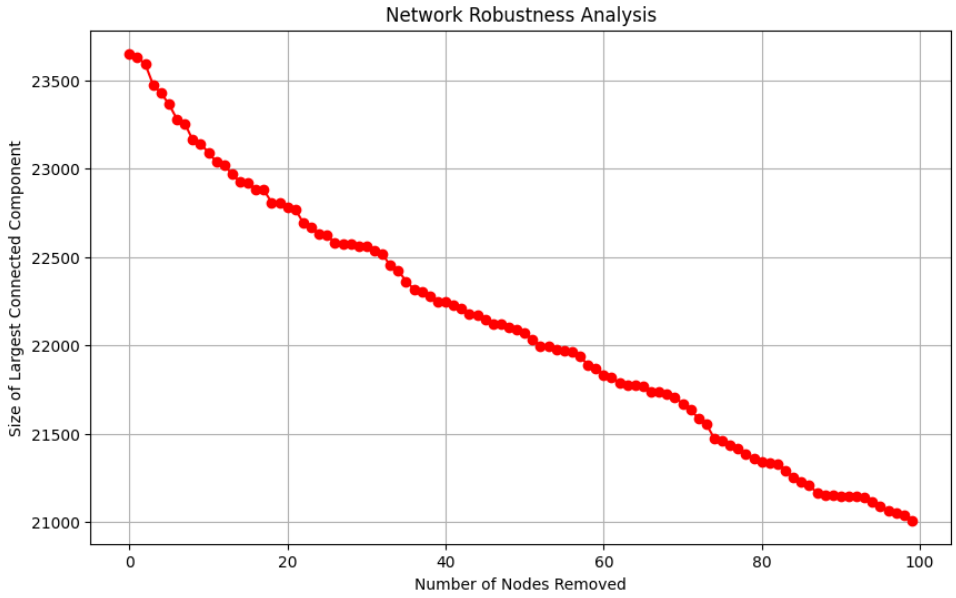

Figure 5: Effect of Node Removal on the Largest Connected Component

5.1.4 Centrality Analysis of Key Nodes

To determine the most influential nodes in the network, we examined four centrality measures—degree, betweenness, closeness, and PageRank—across three node types: papers, organizations, and authors. Each measure provides a distinct perspective on influence and connectivity, offering insights into the structural importance of entities within the academic collaboration network.

5.1.4.1 Degree Centrality

Degree centrality quantifies a paper's influence based on its number of direct connections within the network. Figure 6 highlights the most connected research papers, with *"HyperCLOVA X Technical Report"* and *"Introducing v0.5 of the AI Safety Benchmark from MLCommons"* exhibiting the highest connectivity.

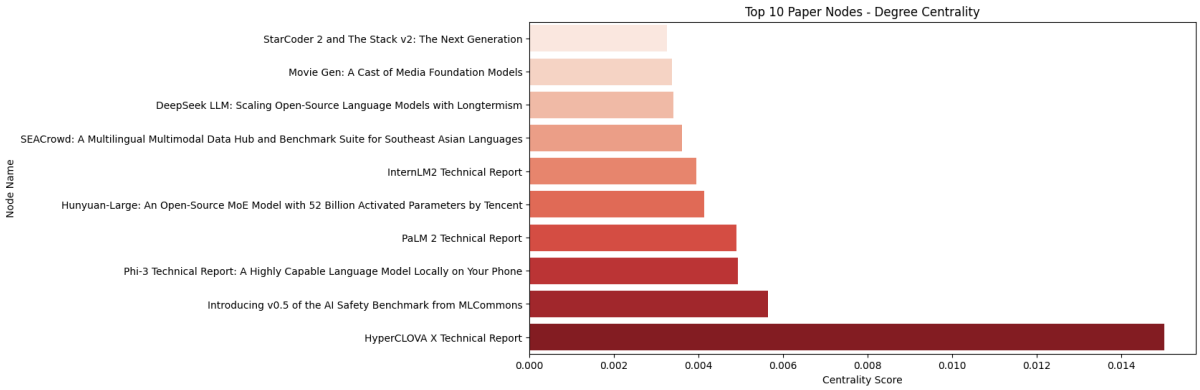

Figure 6: Top 10 papers by Degree Centrality in the Network

5.1.4.2 Betweenness Centrality

Betweenness centrality quantifies a paper's role in facilitating knowledge flow by acting as a bridge between different research actors. Figure 7 highlights papers with the highest betweenness scores. Papers such as *"Introducing v0.5 of the AI Safety Benchmark from MLCommons"* and *"HyperCLOVA X Technical Report"* exhibit the highest centrality. The removal of these papers would significantly impact the network's connectivity.

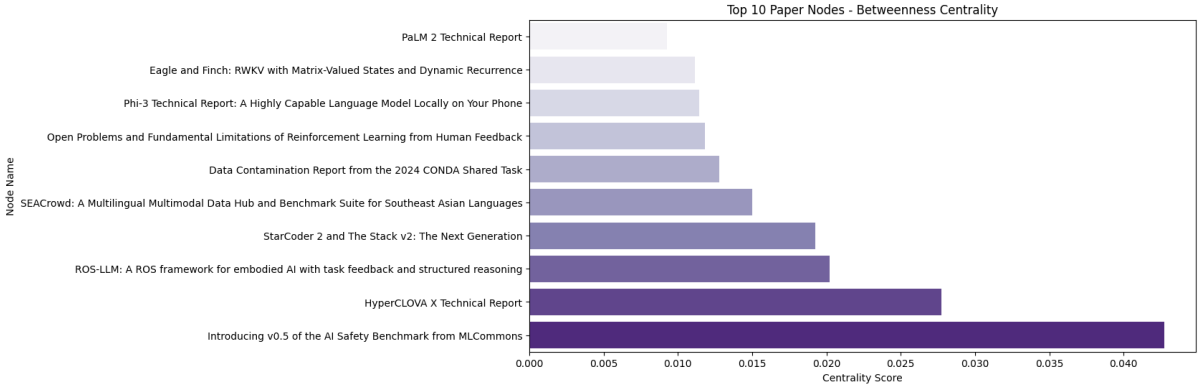

Figure 7: Top 10 Papers by Betweenness Centrality in the Network

5.1.4.3 Closeness Centrality

Closeness centrality measures how efficiently a paper can spread information across the network. Figure 8 highlights key papers with high closeness centrality. Papers such as *"Introducing v0.5 of the AI Safety Benchmark from MLCommons"*, *"ROS-LLM"*, and *"GMAI-VL-5.5M"* exhibit the highest scores, suggesting their crucial role in facilitating the swift exchange of research insights across various researchers.

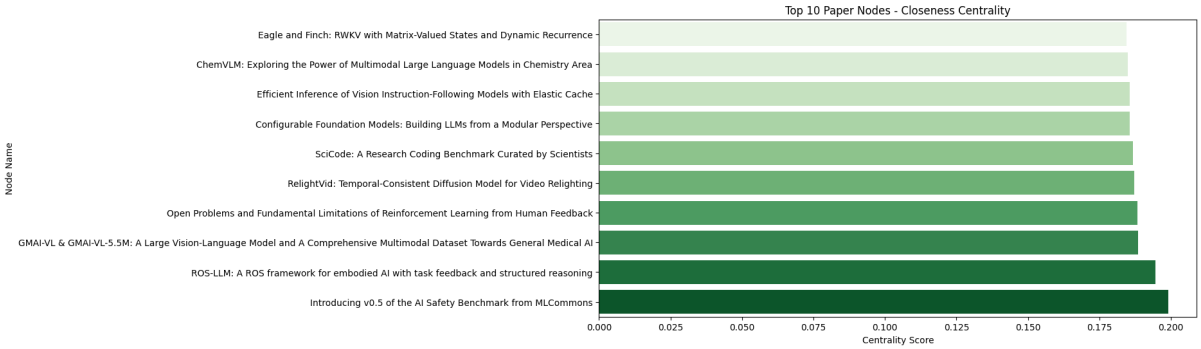

Figure 8: Top 10 papers by Closeness Centrality in the Network

5.1.4.4 PageRank Centrality

PageRank centrality evaluates the influence of papers based on their connections and the importance of the nodes they link to. Figure 9 highlights *HyperCLOVA X Technical Report* and *Introducing v0.5 of the AI Safety Benchmark from MLCommons* as the most authoritative papers in the network. Their high PageRank scores suggest that numerous authors, or organizations of high importance have contributed to their research.

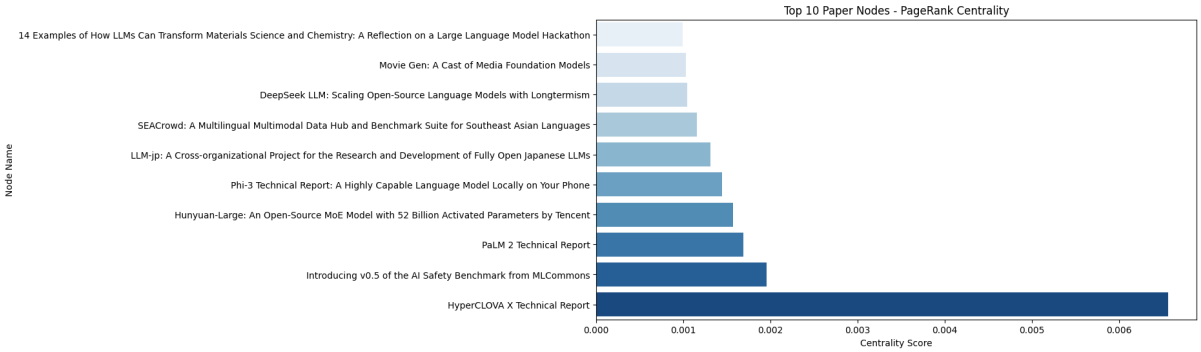

Figure 9: Top 10 papers by PageRank Centrality in the Network

5.2 Co-authorship Network

In this section, we analyze the co-authorship network (Figure 10). It is a weighted network derived from paper-author relationships. Each node in this network represents an author, and edges indicate that two authors co-authored the paper(s), with the weight showing the number of papers they have collaborated with each other. This network reveals how authors collaborate, hinting at the presence of a few highly connected hub authors and multiple specialized communities.

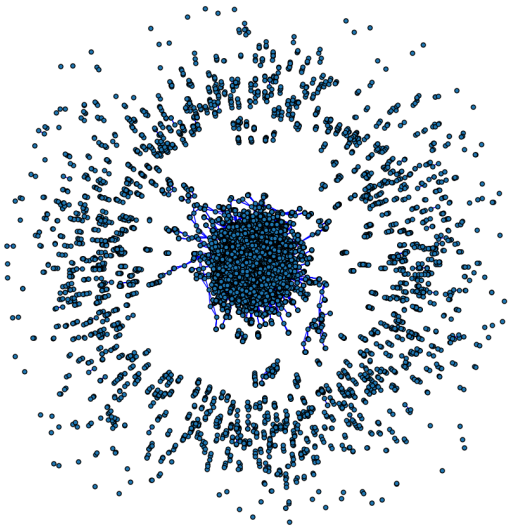

Figure 10: An overall figure of the co-authorship network.

5.2.1 Centrality Measures

To reveal interesting information from this network, we calculate Weighted Degree, Betweenness Centrality, and PageRank. In Figure 11, the Weighted Degree measures show each author's total co-authorship volume (summing all collaboration links), as an example Yu Qiao appears to have the highest total co-authorship count among the authors. The Betweenness Centrality indicates how frequently an author serves as a bridge on the shortest paths between other authors. The PageRank highlights authors whose colleagues are themselves influential.

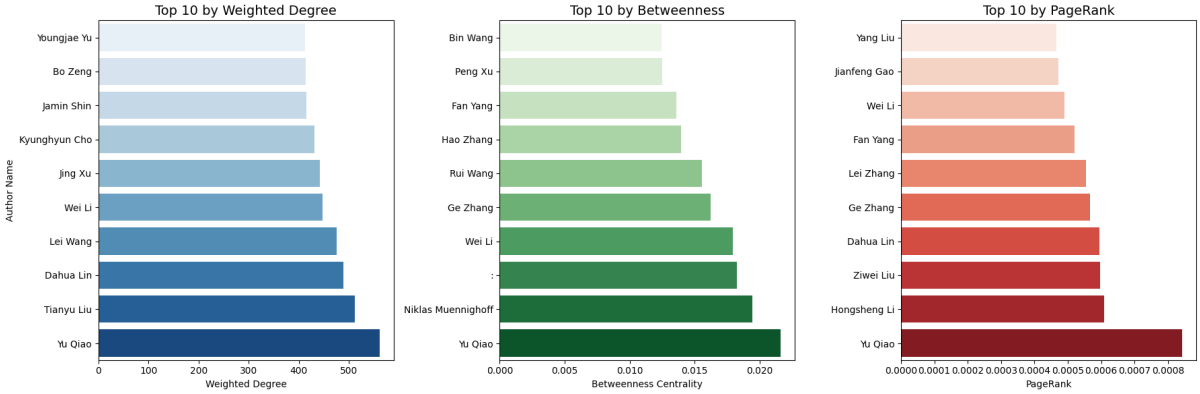

Figure 11: Top 10 authors in a co-authorship network ranked by weighted degree, betweenness centrality, and PageRank.

5.2.2 Modularity

Modularity measures how a network is divided into communities. This network has a strong community structure, with a modularity score of 0.8143. This means authors tend to collaborate more within their groups rather than with others outside their group. Additionally, the network is made up of tightly connected clusters with only a few connections between them.

5.2.3 Community

Our analysis of the co-authorship network uncovered 53 unique communities, showing a diverse and varied structure. The largest of these includes 1.174 authors, making it a major hub for collaboration, likely representing a well-connected research group. This community is made up of 552 organizations, such as *New York University*, *J.P. Morgan AI Research*, *Microsoft*, *Peking University*, and *Georgia Institute of Technology*. On the other hand, the network has many smaller communities (e.g. 5 authors) which might reflect niche research areas, new collaborations, or specialized fields with fewer participants.

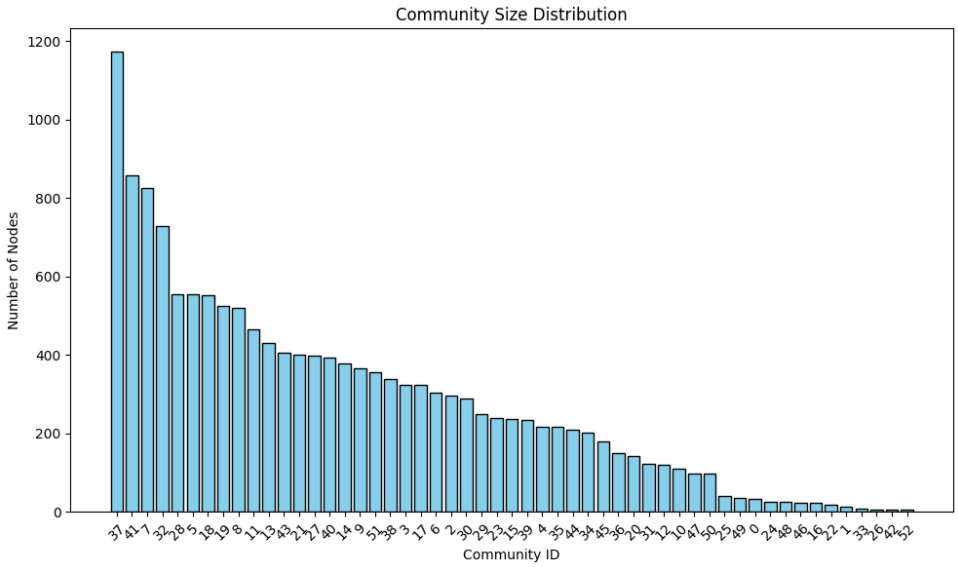

Figure 12: Communities in Co-authorship Network

5.2.4 Global Clustering Coefficient (Transitivity)

The Global Clustering Coefficient is a measure of how tightly connected the nodes in a network are. It quantifies the tendency of nodes to form triangles (groups of three nodes that are all connected). The global clustering coefficient for this network is 0.941, which is very close to the maximum value of 1. This indicates that this network has a highly clustered structure, meaning that authors in the network tend to form tightly connected groups.

5.2.5 Scale-free

The degree distribution of the co-authorship network follows a power-law distribution with an alpha value of 1.87, indicating a scale-free structure, which means that it consists of a few authors that are highly connected, acting as central hubs, while most authors have fewer collaborations. This suggests that the network is shaped by a small number of key players who collaborate extensively, while the majority of authors work with fewer partners.

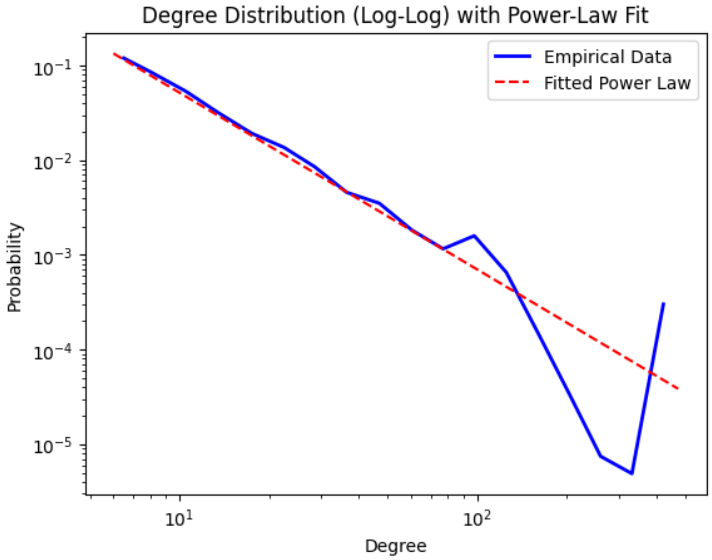

Figure 13: Power-law in Co-authorship Network

5.2.6 Clique

A clique is a group of nodes where every pair of nodes is directly connected, and the maximal cliques are those that cannot be extended by adding any other node from the graph without losing the clique property. The size of the largest clique is 396, which reveals the paper with the most authors since authors of a paper naturally form a clique.

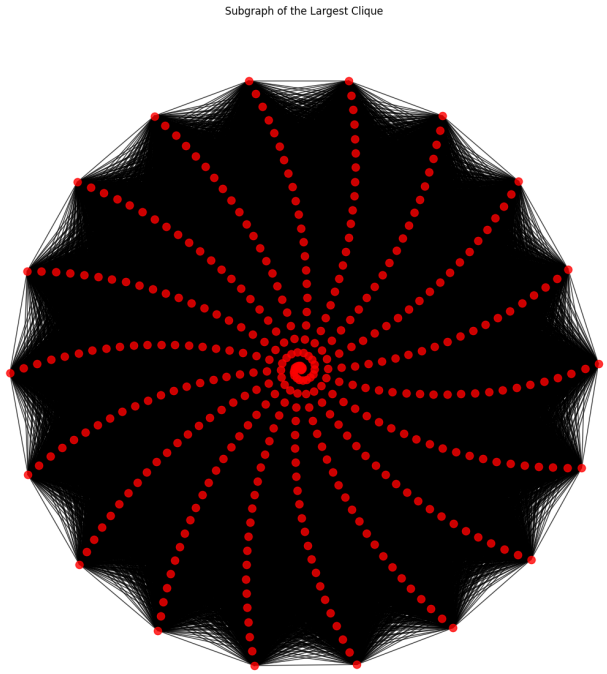

Figure 14: Largest Clique in Co-authorship Network

5.3 Co-organization Network

In this section, we analyze the Co-organization network. It is a weighted network derived from paper-organization relationships. Each node in this network represents an organization, and edges indicate that two organizations have collaborated, with the weight indicating the number of collaborations.

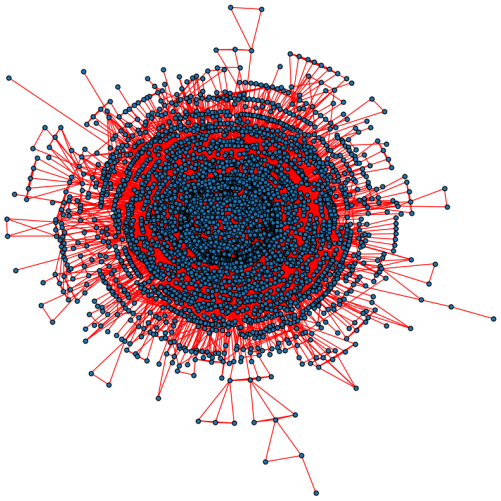

Figure 15: Co-organization network consists of 3669 nodes and 14572 egdes

5.3.1 Centrality Measures

Similar to the co-author network, we calculate Weighted Degree, Betweenness Centrality, and PageRank. In Figure 16, the Weighted Degree metric represents the total collaboration volume of each organization, organizations with high weighted degrees are highly active in collaborative research. The Betweenness Centrality metric measures how often an organization acts as a bridge on the shortest paths between other organizations. Organizations with high betweenness centrality play a critical role in connecting different parts of the network. The PageRank metric identifies organizations whose collaborators are themselves influential. It reflects not only the number of collaborations but also the importance of the organizations they collaborate with.

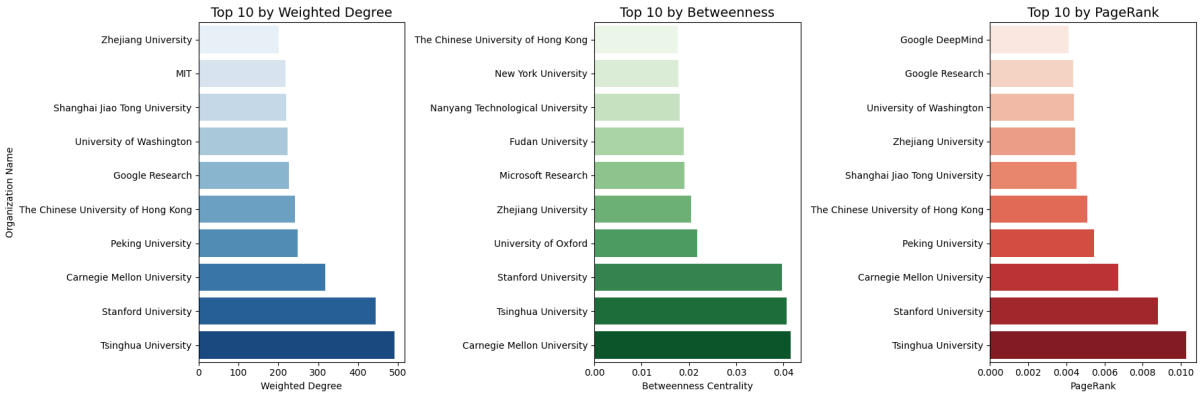

Figure 16: Top 10 organizations in a co-organization network ranked by weighted degree, betweenness centrality, and PageRank.

5.3.2 Modularity

Modularity measures how a network is divided into communities. This network has a strong community structure, with a modularity score of 0.5948. Therefore, the network has distinct and well-defined clusters where organizations within the same cluster collaborate more frequently with each other than with organizations outside their cluster.

5.3.3 Community

Our analysis of the co-organization network uncovered 34 unique communities. The community size varies significantly, ranging from 3 to 620 nodes. By exploring the biggest community, we conclude that *Tsinghua University* consists of 165 papers, and the organization in the smallest community is *DP Technology*, with 2 papers.

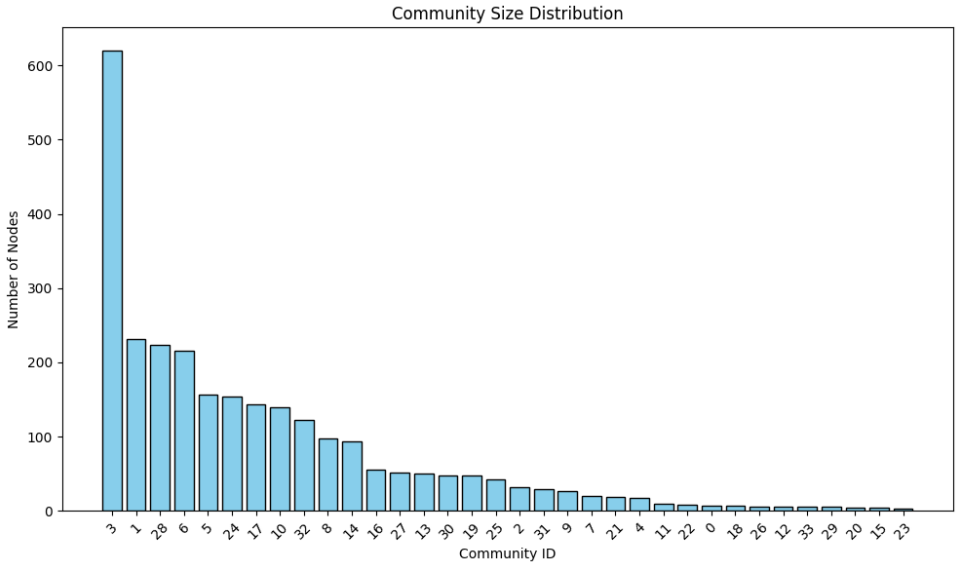

Figure 17: Communities in Co-organization Network

5.3.4 Global Clustering Coefficient (Transitivity)

Global Clustering Coefficient is a tendency for organizations to form tightly-knit collaborative triangles (where two collaborators of an organization are also likely to collaborate). The global clustering coefficient for this network is 0.287, which shows that there is a tendency to form tightly-knit collaborative triangles, the network is not highly clustered.

5.3.5 Scale-free

The analysis of the degree distribution using the power-law reveals that the network exhibits a scale-free property, with an alpha value of approximately 2.08. This alpha value indicates the presence of a few highly connected nodes and many nodes with low degrees, which is typical in scale-free networks.

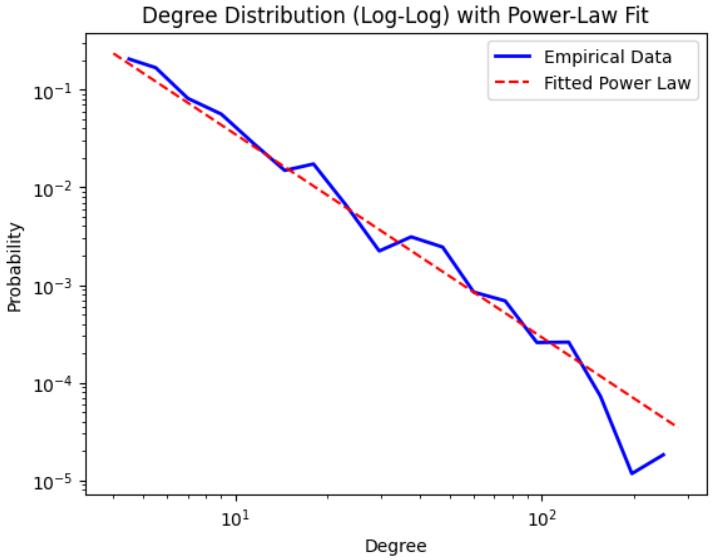

Figure 18: Power-law in Co-organization Network

5.3.6 Clique

The largest clique in this network includes 49 nodes, relating to the paper *Introducing v0.5 of the AI Safety Benchmark from MLCommons*, which has the largest number of organizations involved in the research process.

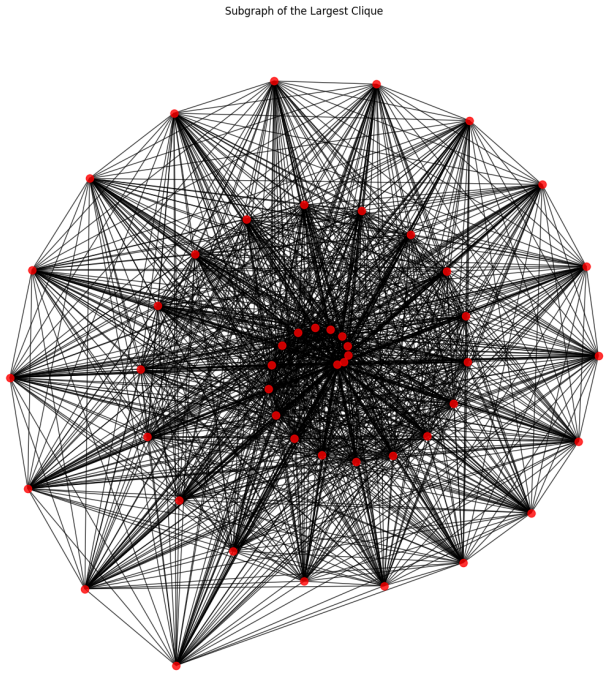

Figure 19: Largest Clique in Co-organization Network

6 Conclusion

This study presents a structured network analysis of modern AI research, focusing on the important papers, organizations, and collaborations within the Hugging Face Daily Papers. By

constructing and examining three distinct networks—the Core Network, Co-authorship Network, and Co-organization Network—we provide a comprehensive view of the AI research landscape.

Our results show that AI research follows a scale-free network pattern, meaning that most papers, authors, and organizations have only a few connections, while a small number are highly connected and play a key role in spreading knowledge.

Centrality analysis identifies the most influential papers, including "*HyperCLOVA X Technical Report*" and "*Introducing v0.5 of the AI Safety Benchmark from MLCommons*". These papers act as reference points in AI research, influencing many different fields. Our analysis of co-authorship and co-organization networks also shows that leading organizations like "*Tsinghua University*" and "*Google Research*" are at the center of AI research collaborations.

Our robustness analysis reveals if highly connected key players are removed, the network breaks apart quickly. This shows that AI research depends on a small number of influential researchers and organizations, making the field somewhat fragile.

Community detection shows that AI research is highly modular, meaning that it is divided into specialized clusters. The largest research community includes 1,174 authors and involves organizations such as "New York University", "J.P. Morgan AI Research", "Microsoft", and "Peking University". This suggests that AI research is divided into many focused areas of study.

Our study answers the research motivation by providing a clear and structured way to track AI research trends. By identifying key contributors and collaboration patterns, we offer useful insights for researchers, industry professionals, and policymakers. These insights help in choosing future research directions, finding collaboration opportunities, and allocating resources wisely.

7 Critique

The study successfully conducts an analysis of the research landscape of the most influential AI papers of the previous years. Our graph dataset, which consists of papers, authors, organization, and their various relationships can be a key feature for other researchers to survey this network from different angles. Additionally, our method of creating the dataset is also useful for those who are interested in analyzing AI papers.

While our effort to create the dataset included overcoming many challenges and limitations imposed by the open APIs, we succeeded in assembling the dataset through multiple steps. One of these steps, extracting the organizations of a paper using LLMs, has the potential to improve, by ensuring the LLM is identifying the organizations with a high accuracy. It can be a further improvement to know what the quantitative accuracy of the LLM is in identifying those organizations, however, this requires labeled datasets of papers and their organizations and remains an area for future improvements. Further, identifying which organizations belong to the academia, and which to the industry can be highly insightful to understand the distribution of model AI research among various entities.

The analysis successfully answers some key questions regarding the current state of AI research, the role of industry and academia in pushing it forward, the concentration or distribution of research among actors, identifying key players, and the quantitative relationships between various entities. However, we believe the dataset has much more room for exploration, defining further questions, and finding their answers.

Overall, this study lays strong foundations for further studies that aim to analyze AI research papers in different contexts, through different metrics, or to answer a new set of questions. This work provides a transparent analysis of various key fundamental questions regarding the current state of AI research, encouraging more researchers to explore this idea or define new ones.

References

[1] Daily Papers - Hugging face. (2001, February 20). https://huggingface.co/papers

[2] Study: Industry now dominates AI research | MIT Sloan. (2023, May 18). MIT Sloan. https://mitsloan.mit.edu/ideas-made-to-matter/study-industry-now-dominates-ai-research

[3] arXiv API Access - arXiv info. (n.d.). https://info.arxiv.org/help/api/index.html

[4] scholarly. (2023, January 16). PyPI. https://pypi.org/project/scholarly/

[5] Google Scholar. (n.d.). https://scholar.google.com/

[6] ORCID. (2024, May 22). Public API - ORCID. https://info.orcid.org/what-is-orcid/services/public-api/

[7] Ollama. (n.d.). Ollama. https://ollama.com/

[8] Team, Q. (2024, September 18). Qwen2.5: A party of foundation models! Qwen. https://qwenlm.github.io/blog/qwen2.5/