A

Project Report On

“Case Study on Global Innovation Network and Analysis (GINA)”

Submitted in partial Fulfilment of Bachelor of Engineering in Computer Engineering,

by Savitribai Phule Pune University Submitted to

Sr.no	Name	PRN.no	SEAT.no
1	CHETAN POPAT DARADE	77721730H	T191174220
2	PRASHANT UDAY BEDADE	72221718J	T191174209

Guided By

Prof. S. S. Hinge

SANGHAVI COLLEGE OFENGINEERING,

NAAC-ACCREDITED INSTITUTE WITH “B+” GRADE

Department of Computer Engineering

Nashik-422202

[2023-2024]

Shree Mahavir
Education Society's
Sanghavi
College of Engineering
Nashik

NAAC-ACCREDITED INSTITUTE
WITH “B+” GRADE

CERTIFICATE

This is to certify that **Mr. Chetan P. Darade, Mr. Prashant U. Bedade** has successfully completed the Mini Project in “**DATA ANALYTICS”** work entitled **“CASE STUDY ON GLOBAL INNOVATION AND NETWORK ANALYSIS (GINA)” s**ubmitted “Project report” during the academic year **2023-2024**, in the partial fulfilment of Bachelor of Computer Engineering by Savitribai Phule Pune university.

Date: / / 2024 **Place: Nashik**

Prof. S. S. Hinge	**Prof. P. Biswas**	**Dr. B. S. Shirole**
(Project Guide)	**(H. O. D)**	**(Principal)**

ABSTRACT

The rapid move of China and India from low-cost producers to innovators has triggered an increasing interest in the globalization of innovation activities and more specifically, the surge of global innovation networks (GINs). However, hitherto most of the literature is either theoretical or based on a handful of cases. We do not know what are the different forms of GINs in which firms participate, both in terms of the various degrees of globalness, innovativeness and networkedness, as well as other key characteristics. In this paper, we propose a firm-based taxonomy of global innovation networks that takes into account these different dimensions. This paper provides empirical evidence about the characteristics of the different variants of global innovation networks, observed in five European countries as well as Brazil, China, India and South Africa. It relies on survey-based firm-level data and provides for the first time a theoretical and empirical overview of the different forms of global innovation networks.

Keywords: Globalization, innovation networks, taxonomy, Europe, South Africa, Brazil, China, India

1 Introduction:

EMC's Global Innovation Network and Analytics (GINA) team is a group of senior technologists located in centers of excellence (COEs) around the world. This team's charter is to engage employees across global COEs to drive innovation, research, and university partnerships. In 2012, a newly hired director wanted to improve these activities and provide a mechanism to track and analyses the related information. In addition, this team wanted to create more robust mechanisms for capturing the results of its informal conversations with other thought leaders within EMC, in academia, or in other organizations, which could later be mined for insights. The GINA team thought its approach would provide a means to share ideas globally and increase knowledge sharing among GINA members who may be separated geographically. It planned to create a data repository containing both structured and unstructured data to accomplish three main goals. Store formal and informal data. Track research from global technologists. Mine the data for patterns and insights to improve the team's operations and strategy. The GINA case study provides an example of how a team applied the Data Analytics Lifecycle to analyses innovation data at EMC. Innovation is typically a difficult concept to measure, and this team wanted to look for ways to use advanced analytical methods to identify key innovators within the company.

2 Phase:

2.1: Discovery

In the GINA project's discovery phase, the team began identifying data sources. Although GINA was a group of technologists skilled in many different aspects of engineering, it had some data and ideas about what it wanted to explore but lacked a formal team that could perform these analytics. After consulting with various experts including Tom Davenport, a noted expert in analytics at Babson College, and Peter Gloor, an expert in collective intelligence and creator of Co IN (Collaborative Innovation Networks) at MIT, the team decided to crowd source the work by seeking volunteers within EMC. Here is a list of how the various roles on the working team were fulfilled. Business User, Project Sponsor, Project Manager: Vice President from Office of the CTO
Business Intelligence Analyst: Representatives from IT Data Engineer and Database Administrator (DBA): Representatives from IT Data Scientist: Distinguished Engineer, who also developed the social graphs shown in the GINA case study

The project sponsor's approach was to leverage social media and blogging to accelerate the collection of innovation and research data worldwide and to motivate teams of "volunteer" data scientists at worldwide locations. Given that he lacked a formal team, he needed to be resourceful about finding people who were both capable and willing to volunteer their time to work on interesting problems. Data scientists tend to be passionate about data, and the project sponsor was able to tap into this passion of highly
talented people to accomplish challenging work in a creative way. The data for the project fell

into two main categories. The first category represented five years of idea submissions from EMC's internal innovation contests, known as the Innovation Roadmap (formerly called the Innovation Showcase). The Innovation Roadmap is a formal, organic innovation process whereby employees from around the globe submit ideas that are then vetted and judged. The best ideas are selected for further incubation. As a result, the data is a mix of structured data, such as idea counts, submission dates, inventor names, and unstructured content, such as the textual descriptions of the ideas themselves. The second category of data encompassed minutes and notes representing innovation and research activity from around the world. This also represented a mix of structured and unstructured data. The structured data included attributes such as dates, names, and geographic locations. The unstructured documents contained the "who, what, when, and where" information that represents rich data about knowledge growth and transfer within the company. This type of information is often stored in business silos that have little to no visibility across disparate research teams.

The 10 main IHs that the GINA team developed were as follows:

IH1: Innovation activity in different geographic regions can be mapped to corporate strategic directions.

IH2: The length of time it takes to deliver ideas decreases when global knowledge transfer occurs as part of the idea delivery process.

IH3: Innovators who participate in global knowledge transfer deliver ideas more quickly than those who do not.

IH4: An idea submission can be analyzed and evaluated for the likelihood of receiving funding.

IH5: Knowledge discovery and growth for a particular topic can be measured and compared across geographic regions.

IH6: Knowledge transfer activity can identify research-specific boundary spanners in disparate regions.

IH7: Strategic corporate themes can be mapped to geographic regions.

IH8: Frequent knowledge expansion and transfer events reduce the time it takes to generate a corporate asset from an idea.

IH9: Lineage maps can reveal when knowledge expansion and transfer did not (or has not) resulted in a corporate asset.
IH10: Emerging research topics can be classified and mapped to specific idolators, innovators, boundary spanners, and assets

The GINA (IHs) can be grouped into two categories:
Descriptive analytics of what is currently happening to spark further creativity, collaboration, and asset generation Predictive analytics to advise executive management of where it should be investing in the future

2.2 Data Preparation

The team partnered with its IT department to set up a new analytics sand box to store and experiment on the data. During the data exploration exercise, the data scientists and data engineers began to notice that certain data needed conditioning and normalization. In addition, the team realized that several missing datasets were critical to testing some of the analytic hypotheses. As the team explored the data, it quickly realized that if it did not have data of sufficient quality or could not get good quality data, it would not be able to perform the subsequent steps in the lifecycle process. As a result, it was important to determine what level of data quality and cleanliness was sufficient for the project being undertaken. In the case of the GINA, the team discovered that many of the names of the researchers and people interacting with the universities were misspelled or had leading and trailing spaces in the datastore. Seemingly small problems such as these in the data had to be addressed in this phase to enable better analysis and data aggregation in subsequent phases.

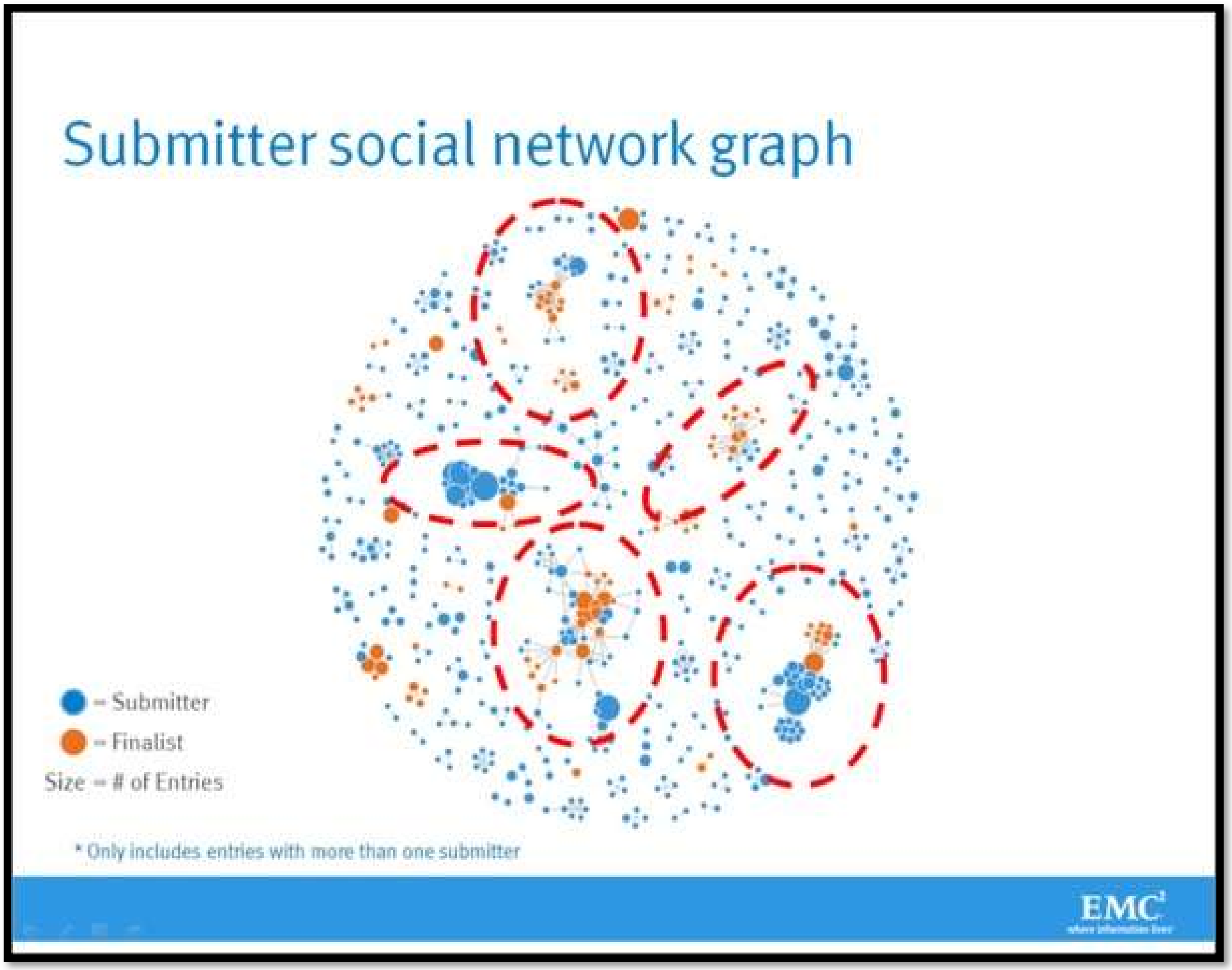
2.3 Model Planning

In the GINA project, for much of the dataset, it seemed feasible to use social network analysis techniques to look at the networks of innovators within EMC. In other cases, it was difficult to come up with appropriate ways to test hypotheses due to the lack of data. In one case (IH9), the team made a decision to initiate a longitudinal study to begin tracking data points over time regarding people developing new intellectual property. This data collection would enable the team to test the following two ideas in the future: IH8: Frequent knowledge expansion and transfer events reduce the amount of time it takes to generate a corporate asset from an idea.IH9: Lineage maps can reveal when knowledge expansion and transfer did not (or has not) result(ed) in a corporate asset. For the longitudinal study being proposed, the team needed to establish goal criteria for the study. Specifically, it needed to determine the end goal of a successful idea that had traversed the entire journey. The parameters related to the scope of the study included the following considerations:

- Identify the right milestones to achieve this goal
- Trace how people move ideas from each milestone toward the goal.
- Once this is done, trace ideas that die, and trace others that reach the goal. Compare the journeys of ideas that make it and those that do not.
- Compare the times and the outcomes using a few different methods (depending on how the data is collected and assembled). These could be as simple as t-tests or perhaps involve different types of classification algorithms.

2.4 Model Building

In Phase 4, the GINA team employed several analytical methods. This included work by the data scientist using Natural Language Processing (NLP) techniques on the textual descriptions of the Innovation Roadmap ideas. In addition, he conducted social network analysis using R and RStudio, and then he developed social graphs and visualizations of the network of communications related to innovation using R'sggplot2 package. Examples of this work are shown in Figures 1.4.1 and 1.4.2.



Submitter social network graph

● = Submitter

● = Finalist

Size = # of Entries

* Only includes entries with more than one submitter

Figure 1 4.1 Social graphs Visualization of idea submitters and finalists

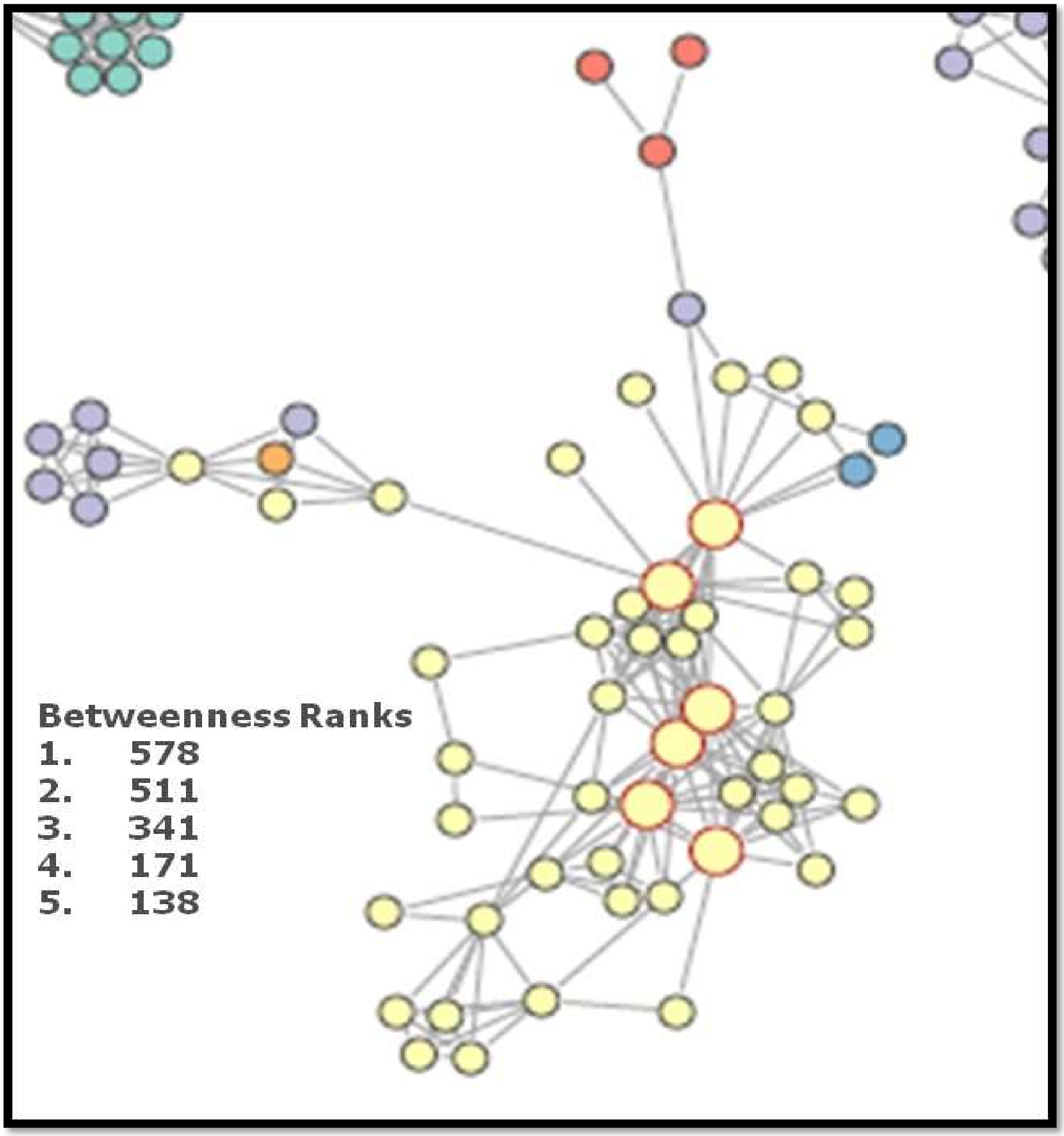


Figure 1.4.2 Social graph visualization of top innovation influencers

Figure 1.4.1 shows social graphs that portray the relationships between idea submitters within GINA. Each color represents an innovator from a different country. The large dots with red circles around them represent hubs. A Huber presents a person with high connectivity and a high “betweenness” score. The cluster in Figure 1.4.2 contains geographic variety, which is critical to prove the hypothesis about geographic boundary spanners. One person in this graph has an unusually high score when compared to the rest of the nodes in the graph. The data scientist identified this person and ran a query against his name within the analytic sand box. These actions yielded the following information about this research scientist (from the social graph), which illustrated how influential he was within his business unit and across many other areas of the company worldwide: In 2011, he attended the ACM SIGMOD conference, which is a top-tier conference on large-scale data management problems and databases. He visited employees in France who are part of the business unit for EMC's content management teams within Documented (now part of the Information Intelligence Group, or IIG). He presented his thoughts on the SIGMOD conference at a virtual brownbag session attended by three employees in Russia, one employee in Cairo, one employee in Ireland, one employee in India, three employees in the United States, and one employee in Israel. In 2012, he attended the SDM 2012 conference in California. On the same trip he visited innovators and researchers at EMC federated companies, Pivotal and VMware. Later on that trip he stood before an internal council of technology leaders and introduced two of his researchers to dozens of corporate innovators and researchers.

This finding suggests that at least part of the initial hypothesis is correct; the data can identify innovators who span different geographies and business units. The team used Tableau software for data visualization and exploration and used the Pivotal Greenplum database as the main data repository and analytics engine.

2.5 Results and Key Findings:

In Phase 5, the team found several ways to cull results of the analysis and identify the most impactful and relevant findings. This project was considered successful in identifying boundary spanners and hidden innovators. As a result, the CTO office launched longitudinal studies to begin data collection efforts and track innovation results over longer periods of time. The GINA project promoted knowledge sharing related to innovation and researchers spanning multiple areas within the company and outside of it. GINA also enabled EMC to cultivate additional intellectual property that led to additional research topics and provided opportunities to forge relationships with universities for joint academic research in the fields of Data Science and Big Data. In addition, the project was accomplished with a limited budget, leveraging a volunteer force of highly skilled and distinguished engineers and data scientists. One of the key findings from the project is that there was a disproportionately high density of innovators in Cork, Ireland. Each year, EMC hosts an innovation contest, open to employees to submit innovation ideas that would drive new value for the company. When looking at the data in 2011, 15% of the finalists and 15% of the winners were from Ireland. These are unusually high numbers, given the relative size of the Cork COE compared to other larger centers in other parts of the world. After further research, it was learned that the COE in Cork, Ireland had received focused training in innovation from an external consultant, which was proving effective. The Cork COE came up with more innovation ideas, and better ones, than it had in the past, and it was making larger contributions to innovation at EMC. It would have been difficult, if not impossible, to identify this cluster of innovators through traditional methods or even anecdotal, word-of mouth feedback. Applying social network analysis enabled the team to find a pocket of people within EMC who were making disproportionately strong contributions. These findings were shared internally through presentations and conferences and promoted through social media and blogs.

2.6 Operationalize

Running analytics against a sandbox Filled with notes, minutes, and presentations from innovation activities yielded great insights into EMC's innovation culture. Key findings from the project include these:

The CTO office and GINA need more data in the future, including a marketing initiative to convince people to inform the global community on their innovation/research activities. Some of the data is sensitive, and the team needs to consider security and privacy related to the data, such as who can run the models and see the results. In addition to running models, a parallel initiative needs to be created to improve basic Business Intelligence activities, such as dashboards, reporting, and queries on research activities worldwide. A mechanism is needed to continually re-evaluate the model after deployment. Assessing the benefits is one of the main goals of this stage, as is defining a process to retrain the model as needed. In addition to the actions and findings listed, the team demonstrated how analytic scan drive new insights in projects that are traditionally difficult to measure and quantify. This project informed investment decisions in university research projects by the CTO office and identified hidden, high-value innovators. In addition, the CTO office developed tools to help submitters improve ideas using topic modelling as part of new recommender systems to help idea submitters find similar ideas and refine their proposals for new intellectual property. Table 1.6.1 outlines an analytics plan for the GINA case study example. Although this project shows only three findings, there were many more. For instance, perhaps the biggest overarching result from this project is that it demonstrated, in a concrete way, that analytics can drive new insights in projects that deal with topics that may seem difficult to measure, such as innovation. Innovation is an idea that every company wants to promote, but it can be difficult to measure innovation or identify ways to increase innovation. This project explored this issue from the standpoint of evaluating informal social networks to identify boundary spanners and influential people within innovation sub networks. In essence, this project took a seemingly nebulous problem and applied advanced analytical methods to tease out answers using an objective, fact-based approach.

Table 1.6.1 Analytic Plan from the EMC GINA Project

Component	GINA Case Study Description
1. Discovery of Business Problem Framed	Identify the overarching business challenge: Enhancing innovation capabilities for sustained competitiveness in the technology sector.
2. Data	Gather diverse datasets from internal and external sources including historical project data, market analyses, customer feedback, industry reports, patent databases, social media analytics, and expert opinions.
3. Model Planning and Analytic Technique	Utilize advanced analytical techniques including text mining, natural language processing (NLP), predictive analytics, network analysis, and machine learning.
4. Results and Key Findings	Uncover emerging technological paradigms, identify regional innovation hotspots, gain competitor intelligence, and distill customer-centric insights.

Another outcome from the project included the need to supplement analytics with a separate data store for Business Intelligence reporting, accessible to search innovation / research initiatives. Aside from supporting decision making, this will provide a mechanism to be informed on discussions and research happening worldwide among team members in disparate locations. Finally, it highlighted the value that can be gleaned through data and subsequent analysis. Therefore, the need was identified to start formal marketing programs to convince people to submit (or inform) the global community on their innovation/research activities. The knowledge sharing was critical. Without it, GINA would not have been able to perform the analysis and identify the hidden innovators within the company

3 Summary:

This case study described the Data Analytics Lifecycle, which is an approach to managing and executing analytical projects. This approach describes the process in six phases.

1. Discovery
2. Data preparation
3. Model planning
4. Model building
5. Results and Key Findings
6. Operationalize

Through these steps, data science teams can identify problems and perform rigorous investigation of the datasets needed for in-depth analysis. In addition, this case Study discussed the seven roles needed for a data science team. Itis critical that organizations recognize that Data Science is a team effort, and a balance of skills is needed to be successful in tackling Big Data projects and other complex projects involving data analytics.

4 Conclusion:

The GINA case study provides an example of how a team applied the Data Analytics Lifecycle to analyses innovation data at EMC. In addition, this case Study discussed the seven roles needed for a data science team. Innovation is typically difficult concept to measure, and this team wanted to look for ways to use advanced analytical methods to identify key innovators within the company. The GINA team thought its approach would provide a means to share ideas globally and increase knowledge sharing among GINA members who may be separated geographically. Store formal and informal data, Track research from global technologists, Mine the data for patterns and insights to improve the team's operations and strategy

5 References:

- https://www.slideshare.net/SurakshaSanghavi/case-study-107609829
- https://bhavanakhivsara.files.wordpress.com/2018/06/data-science-and-big-data analy-nieizv_book.pd
- https://www.studocu.com/
- https://www.studocu.com/in/document/dr-dy-patil-institute-of-engineering-management-and-research/computer-engg/project-002-gina-assignments/27472460
- https://chatgpt.com/
- https://google.com/