

Estudo Original a Ser Replicado

O estudo original que será replicado é intitulado "Um Estudo Bibliométrico sobre a pesquisa em Inteligência Artificial no Brasil", publicado por Luciana Castro Groenner et al., no *Brazilian Journal of Information Science: Research Trends*, volume 16, publicação contínua, 2022. Este estudo teve como objetivo realizar um levantamento da pesquisa em Inteligência Artificial (IA) no Brasil, utilizando uma expressão de busca abrangente para obter artigos relacionados no período de 2011 a 2020 na base de dados Web of Science. A partir de métodos bibliométricos, os registros encontrados foram analisados utilizando os softwares VantagePoint, VOSViewer e Excel.

ARTIGO DE REPRODUÇÃO: UM ESTUDO BIBLIOMÉTRICO SOBRE A PESQUISA EM LARGE LANGUAGE MODELS

Suelio Matias Gomes Sales

RESUMO

Este estudo teve como objetivo realizar uma análise bibliométrica da pesquisa sobre *Large Language Models* (LLMs) no mundo, utilizando dados da base *Web of Science* no período de 2014 a 2024. A partir de métodos bibliométricos, foram analisados 4.450 artigos utilizando os softwares VOSViewer e Python. Este estudo replicou e expandiu a análise original intitulada "Um Estudo Bibliométrico sobre a pesquisa em Inteligência Artificial no Brasil", publicado por Luciana Castro Groenner et al., que investigou a produção científica em IA no Brasil entre 2011 e 2020. As principais mudanças em relação ao estudo original incluem a ampliação do escopo geográfico para uma análise global, a atualização do período de estudo para 2014 a 2024, a alteração da expressão de busca e a inclusão de novas ferramentas analíticas como o Jupyter e a linguagem de programação Python, além do uso do VOSViewer. Os resultados indicam um crescimento exponencial na produção científica sobre LLMs a partir de 2020, destacando-se as publicações em periódicos como IEEE Access e Cureus Journal of Medical Science.

Palavras-chave: Large Language Models; LLM; Inteligência Artificial; Bibliometria.

ABSTRACT

This study aimed to conduct a bibliometric analysis of research on Large Language Models (LLMs) worldwide, using data from the Web of Science database for the period 2014 to 2024. Using bibliometric methods, 4,450 articles were analyzed utilizing the VOSViewer and Python software. This study replicated and expanded the original analysis titled "A Bibliometric Study on Artificial Intelligence Research in Brazil," published by Luciana Castro Groenner et al., which investigated scientific production in AI in Brazil between 2011 and 2020. The main changes compared to the original study include expanding the geographical scope to a global analysis, updating the study period to 2014 to 2024, altering the search expression, and incorporating new analytical tools such as Jupyter and the Python programming language, in addition to using VOSViewer. The results indicate exponential growth in scientific production

on LLMs since 2020, with notable publications in journals such as IEEE Access and Cureus Journal of Medical Science.

Keywords: Large Language Models; LLM; Artificial Intelligence; Bibliometrics.

1 INTRODUÇÃO

O fascínio pelas máquinas e relatos sobre autômatos tem sido comum ao longo da história, desde os desenhos de robôs na forma de cavaleiros medievais por Leonardo Da Vinci em 1495. O primeiro artigo científico sobre inteligência artificial foi publicado em 1943 no Bulletin of Mathematical Biophysics, intitulado "A logical calculus of the ideas immanent in nervous activity" por McCulloch e Pitts, que propuseram modelos simples de neurônios artificiais capazes de realizar operações computacionais através de redes de neurônios interconectados (Russell & Norvig, 2004; Nilsson, 2009). Em 1948, Turing sugeriu a investigação da possibilidade de máquinas exibirem comportamento inteligente no relatório "*Intelligent Machinery*".

Ao longo dos anos, a inteligência artificial (IA) evoluiu significativamente, abrangendo uma variedade de definições e aplicações em diferentes áreas do conhecimento. Segundo Nilsson, a inteligência artificial é “aquela atividade dedicada a tornar as máquinas inteligentes, e inteligência é a qualidade que permite a uma entidade funcionar de maneira adequada e com visão de futuro em seu ambiente” (Nilsson, 2009, p.13).

A área de inteligência artificial tem passado por ciclos de entusiasmo e estagnação. Atualmente, estamos em uma fase de expansão, estimulada por avanços como o lançamento do ChatGPT e pela concorrência tecnológica entre os Estados Unidos e a China (Lee, 2019). Recentemente, o desenvolvimento dos *Large Language Models* (LLMs) tem revolucionado o campo do processamento de linguagem natural (PLN) e da inteligência artificial, trazendo avanços significativos em diversas tarefas, desde tradução automática até geração de texto (Duque-Pereira, Ives & Moura, 2023; Mujadia et al., 2024; OpenAI, 2024).

Com o crescimento das diversas fontes de dados e estudos relacionados ao tema, torna-se essencial selecionar dados que possam gerar informações relevantes e trazer soluções inovadoras para diversos setores (Chen & Zhang, 2014). A proposta deste trabalho é realizar um estudo bibliométrico que visa investigar a pesquisa em LLMs, através da análise de artigos

publicados nos últimos 10 anos na base *Web of Science*, utilizaremos métodos bibliométricos e ferramentas como VOSViewer e Python para mapear a abordagem sobre o tema no cenário global.

2 METODOLOGIA

Este estudo utilizou métodos bibliométricos para avaliar a produção científica e tecnológica relacionada à LLMs. A bibliometria, que aplica métodos estatísticos para analisar publicações acadêmicas, permite examinar a produção de artigos em uma área específica, mapeando unidades acadêmicas e redes de pesquisadores, e gerando indicadores de produção para mensurar a evolução de um tema científico.

Para este trabalho, foi escolhida a base de dados *Web of Science* (WoS) devido à sua abrangência e relevância internacional, além de ser amplamente utilizada em estudos bibliométricos. A WoS oferece uma coleção multidisciplinar que inclui periódicos de alta relevância, sendo a base para o cálculo do JCR Impact Factor¹, o que garante a representatividade da amostra.

Nos procedimentos metodológicos a coleta de dados foi realizada na WoS. A busca foi estruturada utilizando a palavra-chave "*Large Language Model**" para garantir a inclusão de artigos relevantes sobre o tema e foram encontrados 4.450 artigos que compuseram a base de dados para a análise exploratória. Na etapa de análise bibliométrica, os dados bibliográficos dos artigos foram armazenados em um computador local para tratamento e elaboração de indicadores e os registros foram importados para tratamento e análise preliminar.

A seguir, foram utilizadas ferramentas como o VOSViewer que foi empregado para criar e visualizar redes de colaboração entre autores e instituições, bem como coocorrência de palavras-chave e o Python que foi utilizado para análises estatísticas adicionais e visualizações gráficas. Um ponto importante é que todos os dados e códigos utilizados na análise serão disponibilizados em um repositório no GitHub² para acesso público e transparência.

A partir disso, os seguintes indicadores foram elaborados considerando o período dos últimos 10 anos: 1) Número de publicações indexadas sobre LLMs na WoS, por país; 2)

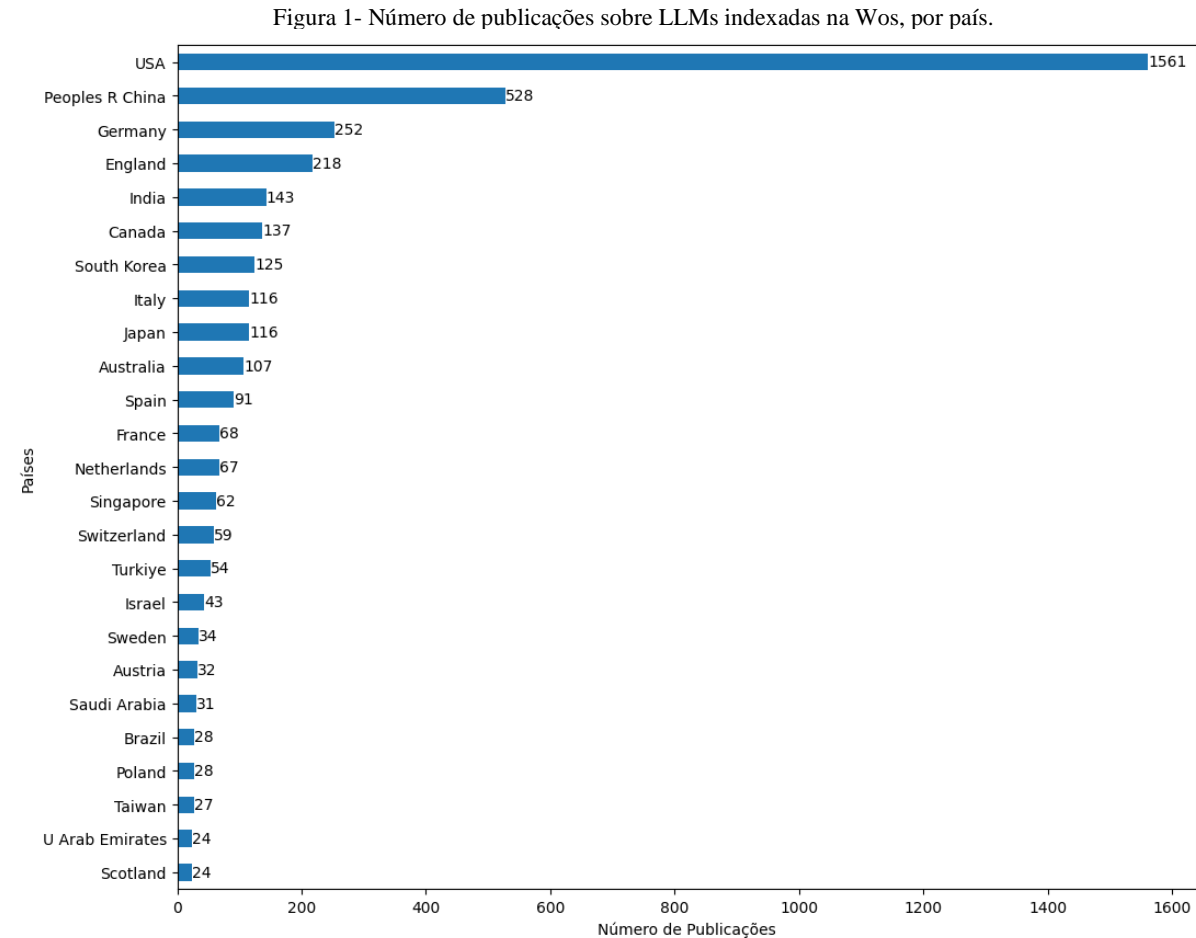
¹ JCR Impact Factor (Fator de Impacto JCR) é uma métrica anual do Journal Citation Reports (JCR) da Clarivate Analytics, que avalia a influência de periódicos científicos.

² Disponível em: https://github.com/sueliomatias/reproducao_fpccII

Crescimento anual do número de publicações indexadas na WoS, por país; 3) Número de publicações indexadas na WoS, por autor; 4) As 20 palavras-chave dos autores de maior incidência em publicações indexadas na WoS.; 5) Áreas de pesquisa com maior número de publicações indexadas na WoS; 6) Rede de coocorrência das 100 palavras-chave de maior ocorrência nas publicações indexadas na WoS; 7) Rede de colaboração entre Instituições com maior número de publicações indexadas na WoS;

3 RESULTADOS E DISCUSSÕES

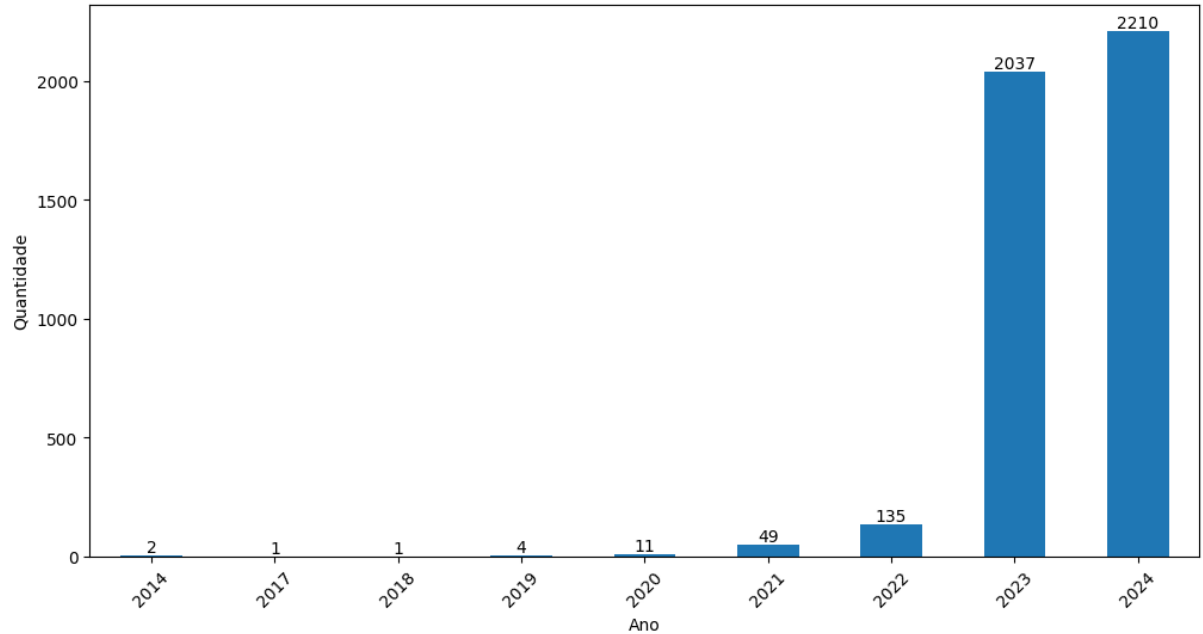
Como primeiro resultado é possível observar na Figura 1 os dados das publicações indexadas na Web of Science (WoS) destacam os Estados Unidos (1561 artigos) e a China (528 artigos) como os principais contribuintes à produção científica global. Na Europa, Alemanha (252), Inglaterra (218), Itália (116) e França (68) se destacam, refletindo seus fortes investimentos em pesquisa e desenvolvimento. Destaque também para o Canadá (137) na América do Norte.



A Ásia também tem mostrado um crescimento significativo, com a Índia (143), Coreia do Sul (125), Japão (116) e Singapura (62) contribuindo de maneira notável indicando uma expansão da produção científica na região. A presença de publicações de países de todos os continentes, como Austrália (107) e Brasil (28), revela um engajamento global na produção científica.

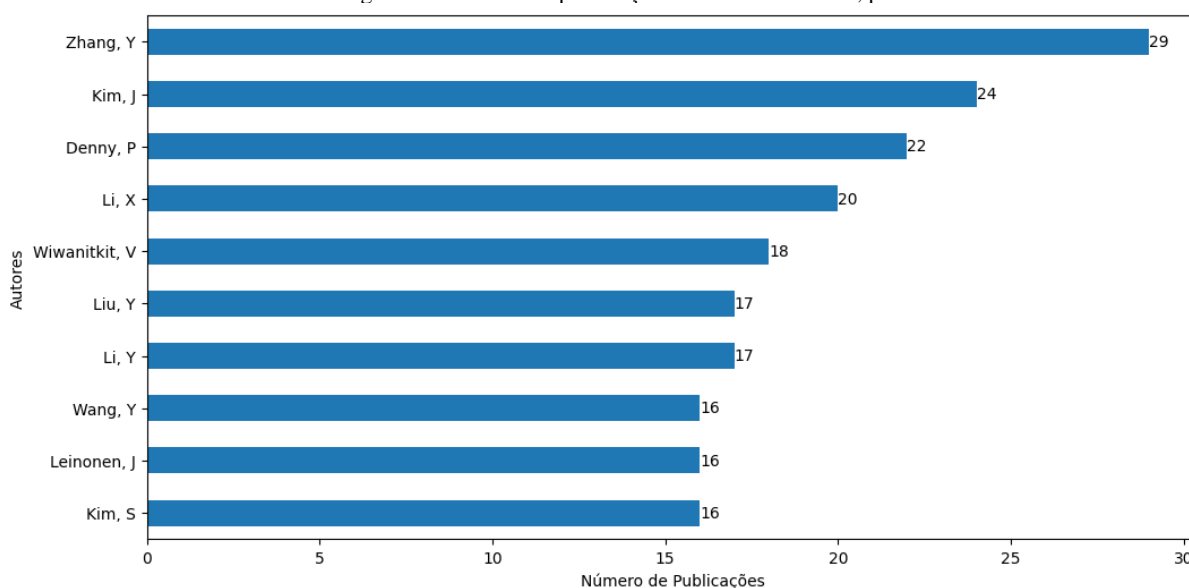
A análise do número de publicações por ano (Figura 2) revela um crescimento significativo na produção científica sobre *Large Language Models* (LLMs) a partir de 2020. Este aumento acentuado coincide com os avanços tecnológicos e a popularização de LLMs, como o GPT-3, que despertaram grande interesse na comunidade científica e em diversas indústrias. Em 2023, foram publicadas 2.037 pesquisas, e o ano de 2024 já apresenta um número considerável de 2.210 publicações, indicando uma tendência de crescimento contínuo.

Figura 2 – Crescimento anual do número de publicações indexadas na Wos, por país e mundo.



Os autores mais produtivos na pesquisa sobre LLMs estão listados na Figura 3. Zhang, Y lidera com 29 publicações, seguido por Kim, J (24 publicações) e Denny, P (22 publicações). Outros autores notáveis incluem Li, X (20 publicações) e Wiwanitkit, V (18 publicações). A diversidade dos autores mais produtivos reflete a ampla colaboração internacional e a contribuição de pesquisadores de várias regiões do mundo.

Figura 3 – Número de publicações indexadas na WoS, por autor.



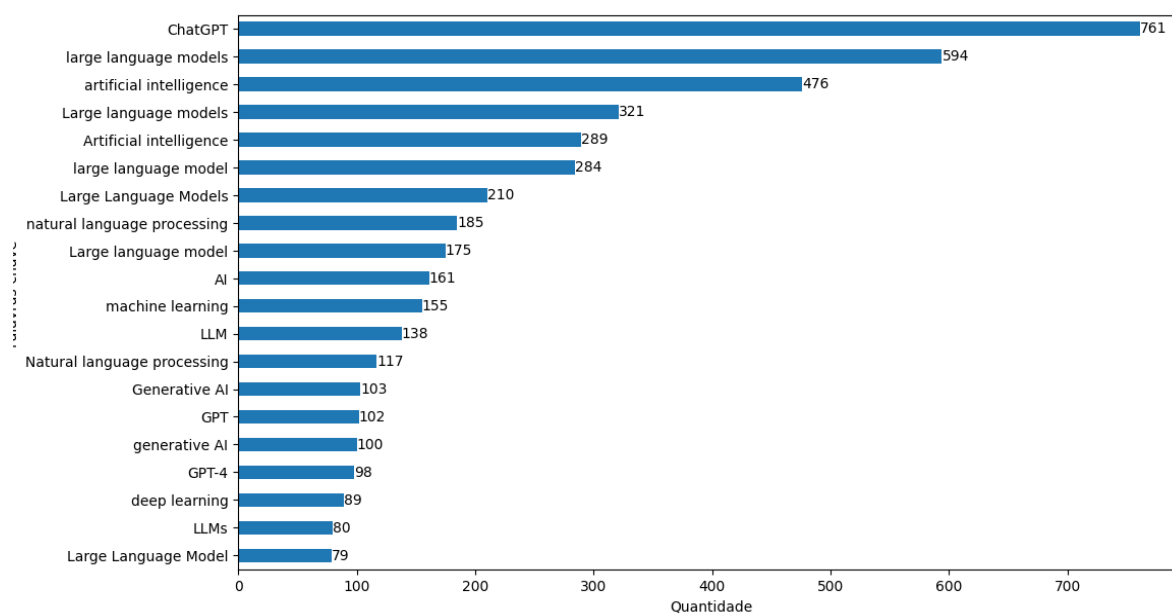
A análise das 20 palavras-chave mais frequentes nas publicações sobre LLMs indexadas na Web of Science revela os principais temas e áreas de pesquisa que têm sido explorados. A Figura 4 apresenta essas palavras-chave, destacando as áreas de maior incidência.

A palavra-chave "ChatGPT" é a mais frequente, com 761 ocorrências, seguida por "large language models" com 594 ocorrências. Isso indica que uma parte significativa da pesquisa recente está focada em tecnologias específicas de LLMs desenvolvidas por organizações como a OpenAI. Termos variáveis para "Large Language Models" e "Artificial Intelligence" aparecem várias vezes na lista, refletindo a diversidade de nomenclaturas usadas pelos autores, mas todas referindo-se a conceitos centrais de LLMs.

"Natural language processing" (185 ocorrências) é outra palavra-chave destacada, indicando que muitos estudos sobre LLMs estão centrados em aplicações de processamento de linguagem natural (NLP). Esta área inclui tarefas como tradução automática, geração de texto e compreensão de linguagem natural. A presença de "machine learning" (155 ocorrências) e "deep learning" (89 ocorrências) reforça a importância das técnicas de aprendizado de máquina e redes neurais profundas no desenvolvimento e aplicação de LLMs.

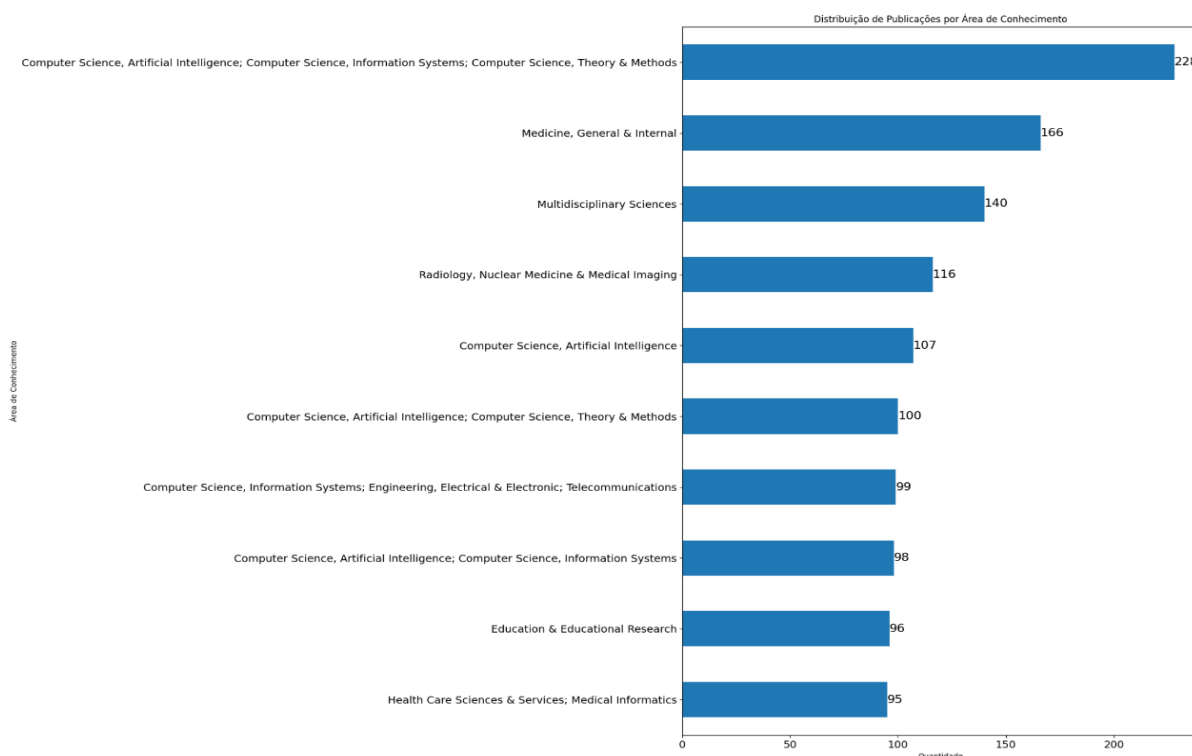
"Generative AI" (103 ocorrências) e "GPT" (102 ocorrências) mostram o interesse crescente em modelos generativos de IA, como a série GPT da OpenAI, que tem sido amplamente discutida e aplicada em diversas áreas. A ocorrência de "GPT-4" (98 ocorrências) evidencia a atenção específica dada a versões mais recentes e avançadas desses modelos, sugerindo um foco contínuo na evolução e aprimoramento das capacidades dos LLMs.

Figura 4 – As 20 palavras-chave dos autores de maior incidência em publicações indexadas na WoS.



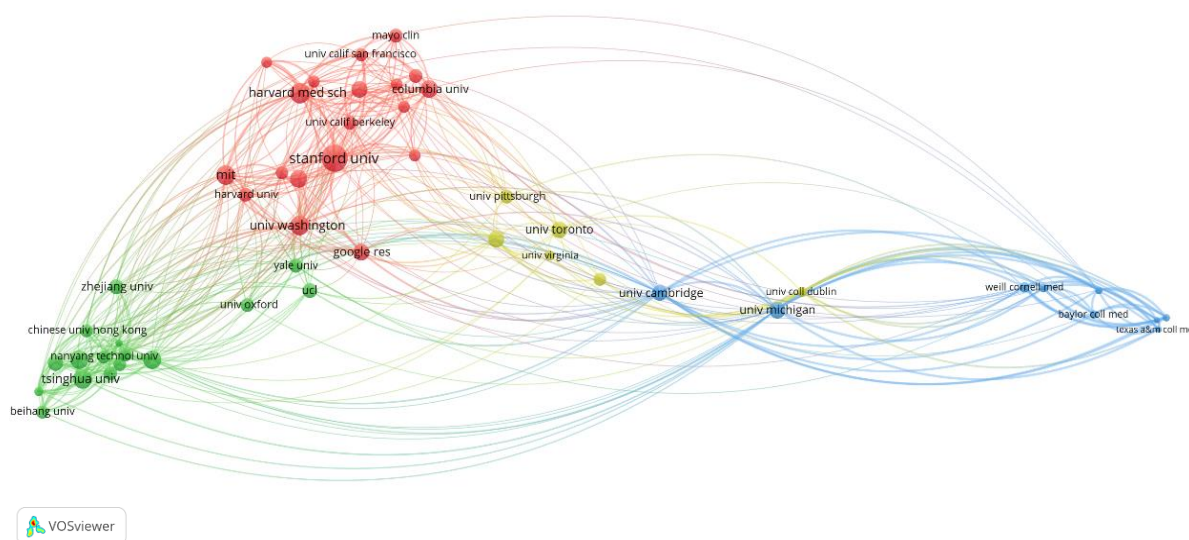
A análise das principais categorias de pesquisa (Figura 5) mostra que a "*Computer Science, Artificial Intelligence; Computer Science, Interdisciplinary Applications*" lidera com 228 publicações, seguida por "*Medicine, General & Internal*" com 166 publicações. Outras categorias notáveis incluem "*Multidisciplinary Sciences*" (140 publicações) e "*Radiology, Nuclear Medicine & Medical Imaging*" (116 publicações). Esses dados indicam que a pesquisa em LLMs é interdisciplinar, abrangendo tanto as ciências da computação quanto a medicina e outras áreas científicas.

Figura 5 – Áreas de pesquisa com maior número de publicações indexadas na WoS.



O papel das universidades e centros de pesquisa é essencial para impulsionar o desenvolvimento tecnológico e social. Em uma análise detalhada, foram identificadas as 50 instituições de ensino e pesquisa com maior contribuição na publicação de trabalhos sobre LLMs em todo o mundo, conforme mostrado na Figura 7. Essas instituições foram responsáveis por um total de 1889 publicações. A Universidade de Stanford destacou-se como a instituição com o maior número de publicações indexadas, refletindo sua liderança e influência no campo dos LLMs. Em seguida, a Harvard Medical School, Universidade de Washington e o MIT aparecem como importantes contribuintes, reforçando a relevância de suas pesquisas na área. Essas instituições não apenas produzem um volume significativo de publicações, mas também desempenham um papel crucial na formação de novas gerações de pesquisadores e na aplicação prática das tecnologias de LLMs.

Figura 7 – Rede de colaboração entre Instituições com maior número de publicações indexadas na WoS



4 CONCLUSÃO

Neste estudo, foi realizada uma análise bibliométrica da produção sobre *Large Language Models* (LLMs) no período de 2014 a 2024. Utilizando uma expressão de busca abrangente, obtivemos um número significativo de publicações na base de dados *Web of Science*, permitindo avaliar o crescimento da produção científica nessa área. Observamos um aumento expressivo nas publicações a partir de 2020, indicando um interesse crescente e um avanço significativo na pesquisa sobre LLMs.

Os dados coletados também destacaram um forte engajamento global na produção científica, com os Estados Unidos e a China liderando o número de publicações. A pesquisa em

LLMs mostrou-se altamente interdisciplinar, abrangendo não apenas as ciências da computação, mas também áreas como medicina e outras ciências.

Apesar do crescimento observado, a produção brasileira ainda é relativamente baixa em comparação com outros países. Os resultados deste estudo podem servir como um ponto de partida para fomentar mais pesquisas sobre LLMs no Brasil, incentivando o desenvolvimento de políticas públicas que promovam a inclusão tecnológica e apoiem o avanço da inteligência artificial no país.

Este trabalho contribuiu para a compreensão do panorama atual da pesquisa em LLMs, evidenciando a necessidade de maior investimento e colaboração interdisciplinar para fortalecer a presença brasileira no cenário global da inteligência artificial.

5 REFERÊNCIA

Artificial Intelligence: A History of Ideas and Achievements. Cambridge: Cambridge University Press, 2009, p. 13. Acesso em: 28 jun. 2024.

CASTRO GROENNER, Luciana et al. Pesquisa em Inteligência Artificial no Brasil.

CHEN, Hsinchun; ZHANG, Yilu. Big Data in Healthcare: A Survey. Journal of Biomedical Informatics, 2014. Acesso em: 29 jun. 2024.

DUQUE-PEREIRA, Ives; MOURA, Sergio. Advances in Large Language Models. Journal of Artificial Intelligence, 2023. Acesso em: 01 jul. 2024.

GIL, Antonio Carlos. Métodos e técnicas de pesquisa social. 5. ed. São Paulo: Atlas, 2002.

HAYASHI, Cristina. Cientometria: Conceitos e aplicações. Revista Brasileira de Biblioteconomia e Documentação, 2013. Acesso em: 02 jul. 2024.

LEE, Kai-Fu. AI Superpowers: China, Silicon Valley, and the New World Order. Boston: Houghton Mifflin Harcourt, 2019. Acesso em: 29 jun. 2024.

MCCULLOCH, Warren S.; PITTS, Walter. A logical calculus of the ideas immanent in nervous activity. Acesso em: 27 jun. 2024.

NILSSON, Nils J. The Quest for RUSSELL, Stuart; NORVIG, Peter. Artificial Intelligence: A Modern Approach. 2nd ed. New Jersey: Prentice Hall, 2004. Acesso em: 02 jul. 2024.

MUJADIA, Salim et al. Recent Developments in Natural Language Processing. Computational Linguistics Journal, 2024. Acesso em: 01 jul. 2024.

XU, Lanling et al. Prompting Large Language Models for Recommender Systems: A Comprehensive Framework and Empirical Analysis. 2024. Disponível em: <https://arxiv.org/abs/2401.04997v1>. Acesso em: 01 jul. 2024.