2024 IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE) | 979-8-3503-7800-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/METROXRAINE62247.2024.10797032

An Overview On Large Language Models Across Key Domains: A Systematic Review

Mattia Bruscia
Department of information engineering
University of Pisa
Pisa, Italy
m.bruscia@studenti.unipi.it
ORCID: 0009-0008-3928-2689

Graziano A. Manduzio
Department of information engineering
University of Pisa
Pisa, Italy
grazianoalfredo.manduzio@phd.unipi.it
ORCID: 0009-0009-7421-5868

Federico A. Galatolo
Department of information engineering
University of Pisa
Pisa, Italy
federico.galatolo@unipi.it
ORCID: 0000-0001-7193-3754

Mario G.C.A. Cimino
Department of information engineering
University of Pisa
Pisa, Italy
mario.cimino@unipi.it
ORCID: 0000-0002-1031-1959

Alberto Greco
Department of information engineering
University of Pisa
Pisa, Italy
alberto.greco@unipi.it
ORCID: 0000-0002-4822-5562

Lorenzo Cominelli
Department of information engineering
University of Pisa
Pisa, Italy
lorenzo.cominelli@unipi.it
ORCID: 0000-0001-8333-7538

Enzo Pasquale Scilingo
Department of information engineering
University of Pisa
Pisa, Italy
enzo.scilingo@unipi.it
ORCID: 0000-0003-2588-4917

***Abstract*—This systematic review explores the applications of Large Language Models (LLMs) across a variety of academic disciplines and professional fields. The analysis is structured through a methodical examination of data derived from the Scopus database over the period from 2017 to 2024. We created both annual and comprehensive datasets of articles and related information based on a generic query, which allowed us to track the development and integration of LLMs into different application fields. To this end, we conceived a dedicated approach that includes an analysis of the trends of subject areas and a Pertinence Analysis (PA) to filter out articles that are not genuinely related to LLMs. Additionally, we performed an annual and overall Terminological Relevance Analysis (TRA) using Machine Learning (ML) techniques, and we examined the yearly trends in research areas containing LLM-related articles by observing the relevance indicators of emerging terms. This extensive investigation highlights how LLMs are increasingly being utilized to improve efficiency, accuracy and productivity, particularly in health and healthcare care, guiding the responsible advancement and application of these technologies in sensitive domains.**

***Index Terms*—large language models, systematic review, Scopus database, trend analysis, metadata analysis, pertinence analysis, terminological relevance, health and healthcare, annual trend, professional fields, machine learning**

I. Introduction

LLMs have become fundamental tools in computational linguistics, leveraging advances in deep learning to process and generate text that mimics human communication. These models have significantly evolved since the introduction of the Transformer architecture, which introduced self-attention mechanisms, enabling models to weigh the importance of words in a sentence regardless of their positional distance [1]. Although at that time this seminal work did not introduce the term *Large Language Model*, this innovation paved the way for transformative models like GPT (*Generative Pre-trained Transformer*) [2], extending their influence beyond traditional natural language processing to fields such as text generation and language translation. Building on the Transformer's capabilities, LLMs have revolutionized how machines understand and generate language, marking a paradigm shift in computational linguistics [3]. As a consequence, in the last few years, this topic reached a notable increase in academic publications and general interest, indicating a growing recognition of their potential. Recent surveys on LLMs highlighted their expansion, explained the evolution of their functioning, how they are trained, and also attempted to classify them into 'families' [4]–[6]. Nonetheless, the integration of LLMs across diverse industries such as healthcare [7], robotics [8], [9], and decision support systems [10] has not yet been sufficiently explored, while the pervasive nature of this technology is evident thanks to many specific examples that are present in the literature. For example, Thirunavukarasu et al. demonstrated the significant capabilities of LLMs in transforming medical practices, highlighting their potential to enhance diagnostic accuracy and personalize patient care [11]. Beyond healthcare, LLMs have found application in several other critical sectors like clinical decision support systems, educational technology, or financial sector, demonstrating their versatility and wide-reaching implications. In robotics, Yamazaki et al. developed a scenario-based dialogue system to enhance android interactions, overcoming the limitations of traditional LLM implementations by segmenting tasks like summarization and response generation for more coherent communication [12]. Similarly, Yoshikawa et al. demonstrated how LLMs can automate complex chemical experiments, translating natural language into robotic commands and improving lab operations for non-experts [13]. In social robotics, Onorati et al. introduced dynamic and personalized interactions by adapting dialogue based on users' social media activities, greatly enhancing user engagement [14]. In another research, Serfaty et al. used nonverbal cues such as eye gaze and facial expressions to improve human-robot interactions, showcasing a zero-shot learning approach for adaptable nonverbal communication [15]. In healthcare, Umerenkov et al. propose a LLM based clinical decision systems by processing extensive patient data to improve diagnostics and treatments [16]. Another significant study by Mehandru et al. evaluated LLMs in "Artificial Intelligence Structured Clinical Examinations" to assess their viability as proactive agents in clinical environments, emphasizing their potential to influence healthcare outcomes [17]. In educational technology, Hosseini et al. explored the integration of ChatGPT in education, revealing uncertainties about its effective application, while Kasneci et al. in another research highlighted how LLMs could personalize learning environments to suit individual student needs [18], [19]. Furthermore, in finance, LLMs like BloombergGPT have shown remarkable capabilities in analyzing financial data and performing complex tasks like sentiment analysis and named entity recognition, demonstrating the extensive utility of these models in the sector [20]. These developments underscore the transformative impact of LLMs and illustrate how they build on and expand the capabilities of earlier architectures, setting new benchmarks in the field of artificial intelligence and Machine Learning (ML). However, this rapid increase in their use complicates the task of pinpointing the specific fields in which they are used. Consequently, unlike other articles that merely define and explain the workings of specific LLMs or their adoption within a single scientific domain, and unlike previous reviews that primarily catalog the technical evolution of LLMs, our systematic review identifies the specific fields that have embraced this transformative technology. An innovative aspect of our study is the use of an LLM to disambiguate and refine the results of our literature search, ensuring greater accuracy and relevance. Therefore, this review aims to trace and understand the real-world deployment of LLMs across a broad spectrum of disciplines in recent years. By highlighting the diverse directions and applications of this technology, we provide an in-depth and comprehensive overview of its current and potential impact, shedding light on how LLMs are revolutionizing various fields and guiding future advancements.

II. Proposed work

A. Creation of the dataset

To facilitate a comprehensive study on the usage of LLMs, data acquisition involved automated downloads from the Scopus database using a Python script, which effectively communicated with Elsevier's Scopus API. Utilizing the Scopus API presents both advantages and disadvantages. On the one hand, the API primarily provides access to article metadata such as titles, abstracts, and keywords, rather than full-text content. This limitation means that, while a substantial amount of data can be retrieved about a wide array of articles, the complete documents are often inaccessible directly through the API. However, this method enables researchers to gather and analyze large datasets efficiently, offering a broad overview of research trends and facilitating comprehensive bibliometric analysis. Using the extensive metadata available, we were able to identify patterns, trends, and connections in a vast number of publications, aligning with our primary objective. Focusing on the Abstract, Title, and Keywords sections of each article ensured a targeted and relevant approach, as these sections typically capture the core themes and focal points of the research. If a term or concept such as 'medical' is mentioned in these sections, it is reasonable to infer its centrality to the content of the article, making this approach practical and justifiable. The following are the steps we took to build the dataset for the analysis. **Configuration**: the script was initialized with the relevant API key and base URLs for fetching data about articles, abstracts, and serial titles. **Query Preparation**: the query has been written to retrieve all documents starting from 2017 that mentioned *Large Language Model*, *LLM* or variations thereof, i.e., `(( TITLE-ABS-KEY ( "large language model" ) OR TITLE-ABS-KEY ( "large language models" ) OR TITLE-ABS-KEY ( "LLM" ) OR TITLE-ABS-KEY ( "LLMs" ) ) AND PUBYEAR > 2016)`. **Iterative Requests**: the script made HTTP GET requests to the Scopus API endpoint, iteratively fetching results paginated by 10 articles at a time. This was to ensure that the API's response limits were respected while capturing all relevant data. **Data Handling**: each article's details, such as title, abstract, keywords, publication year, citation count, and subject areas, were extracted. Subject areas were further fetched from both the article and journal level using specific endpoints provided by the Scopus API. **Data Storage**: all successfully retrieved data were stored in CSV files organized by publication year. The script maintained a progress log to restart the process from the last successful fetch in case of interruptions. Before saving any data, the script checked and created necessary directories based on the year of data retrieval to organize the outputs systematically.

TABLE I
NUMBER OF DOCUMENTS BY YEAR OF PUBLICATION, BEFORE AND AFTER PA.

Year	Before PA	After PA
2017	63	0
2018	71	4
2019	79	5
2020	99	13
2021	151	55
2022	137	67
2023	4308	3731
2024	3117	1961

B. Disambiguating research papers through Pertinence Analysis

A Pertinence Analysis (PA) has been conducted to ensure that our dataset only included articles relevant to our study on LLMs. To this aim, we used an ML-based method to examine each CSV file, year by year. This analysis was performed to separate the relevant from the irrelevant articles based on their content. Each CSV containing the metadata of the articles from various years was passed through an LLM, i.e., Mixtral 8x7B [21]. The LLM was prompted to analyze the abstract, author keywords, and index keywords of each article. The specific prompt used for this analysis was: `Respond with 'TRUE' or 'FALSE' only, without any other comment. Does the following text discuss large language models: 'text_to_analyze'.`
Each field was evaluated three times to enhance the accuracy of the analysis, with the three results averaged to determine the pertinence. An article was classified as relevant if any of the three analyzed fields indicated relevance. Table I shows the number of documents found by year, before and after the process.

C. Analysis of Subject Area Trends Over the Years

The results of the query, before the PA, along with the cleaned dataset, have been used to analyze the trends of various subject areas over the years, as shown in Fig. 1. This involved conducting an analysis of how many articles were labeled year-by-year according to Scopus in various subject areas, and the same was done for the articles considered relevant with the subsequent analysis. Subject areas were grouped as follows: Computer Science and Technology, Engineering and Applied Technology, Biological Sciences and Medicine, Social Sciences and Humanities, Communication, Management Sciences, Applied and Interdisciplinary Sciences.

D. Keywords and Relevant terms extraction

This study employed a systematic approach to extract and analyze key terms from a corpus of collected articles using the Orange Text Mining software [22]. This process is vital for discerning thematic concentrations and tracking terminological evolutions within the domain of LLMs across specified periods. In Fig. 2 we provides a detailed visualization of the entire processing pipeline, showcasing each node involved in our knowledge extraction study. The pipeline is designed to optimize code reuse and allow for the selection of specific parts of the dataset by modifying the filtering phase, making the process streamlined and efficient. Detailed below are the specific steps and corresponding function calls utilized throughout the analysis:

- **Corpus Creation** (`CSV File Import`, `Corpus`): initially, the collected articles were loaded into the Orange environment, enabling further processing.
- **Text Preprocessing** (`Process Text`): the text data underwent preprocessing removing stop words, applying stemming, and lemmatization. These steps are essential for reducing words to their base forms and ensuring that the analysis focuses on the content's essence without common linguistic redundancies.
- **Filtering the data** (`Raw and Columns selections`): The selection of specific rows and columns from the preprocessed dataset enables us to dynamically and easily filter by specific years (row selection) and fields (column selection) without the need to prepare separate corpora or redesign the entire pipeline.
- **Vectorization** (`Bag of Words`): after preprocessing, the text was converted into a vector format using the TF-IDF (Term Frequency-Inverse Document Frequency) technique [23]. This method is very effective because it enhances the importance of unique terms in each document while diminishing the impact of those common across many documents. The TF-IDF score acts as a weighting factor in information retrieval, measuring the relevance of a term relative to a collection of documents. The mathematical representation of TF-IDF is given by:

$$(TF)_{i,j} = \frac{\text{Num. of occurrency of term } i}{\text{Total num. of terms in doc. } j}$$

$$(IDF)_i = \ln\left(\frac{\text{Total num. of docs}}{\text{Num. of docs that contains term } i}\right)$$

$$(TF - IDF)_{i,j} = (TF)_{i,j} \cdot (IDF)_i$$

- **Keyword Extraction** (`Extract Keywords`): subsequently, the keywords were identified and scored based on their TF-IDF values. Keywords with higher scores were deemed more significant and indicative of the core topics discussed in the texts.
- **Keyword Visualization** (`Data Table`): a detailed table was automatically compiled, listing the keywords along with their respective TF-IDF scores for each year and cumulatively for the entire period (Tab. II).

III. RESULTS AND CONCLUSIONS

Important observations are immediately visible from the annual publication numbers reported in Table I. It is evident that after the dataset undergoes disambiguation through pertinence analysis, there are no articles discussing large language models (LLMs) prior to the one considered the

Fig. 1. Trends of Subject Areas before (above) and after (below) Pertinence Analysis (PA).

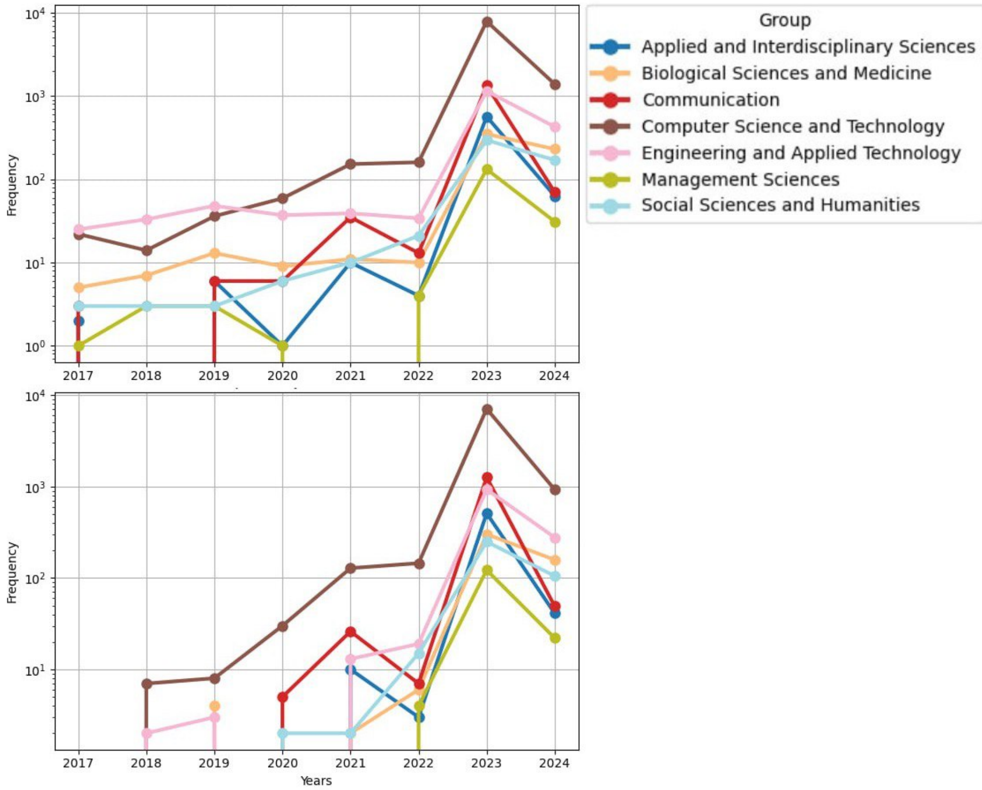

Fig. 2. Detailed pipeline with relative function chain for keyword extraction using Orange [22].

pioneer in this technology [1]. An important observation is that, while LLMs were introduced in 2017 and quickly gained traction within the field of computer science, but it took a few years for their adoption to reach a truly exponential growth, which is identifiable in the year 2023. The analysis of the trends of LLMs divided by different subject areas (Fig. 1), confirmed that, in that year, the explosive growth of LLMs in computer science has finally extended to other sectors, which are now beginning to widely and reliably adopt this technology to enhance diverse solutions. The sector most impacted by this shift is medicine and healthcare, followed by significant increases in engineering, communication, finance, and marketing. This broader adoption reflects the transformative potential of LLM technology across multiple fields. The evolving trends in keyword significance over the years offer a more comprehensive insight into the shifting focus areas within the research and application of LLMs. Our analysis, derived from the frequency and TF-IDF scores of keywords from 2017 to 2024, reveals a gradual yet evident increase in terms not traditionally associated with computer science and NLP. For example, terms like *education*, *medical*, and *human* from 2023 onwards, point to a broadening horizon where LLM applications intersect with diverse disciplines such as education, healthcare, and social sciences. This interdisciplinary trend is not just incidental but reflects a deeper integration of LLM capabilities in addressing complex issues beyond their conventional tech-centric applications. In 2021, terms like *semantic*, *knowledge*, and *understanding* begin to emerge more prominently, suggesting a shift towards more nuanced and context-aware applications of LLMs. By 2024, the persistence of *education* and *medical* alongside *performance* and *information* underscores the role of LLMs in educational frameworks and medical diagnostics, illustrating their penetration into sectors requiring nuanced human-like understanding and decision-making processes. Given these results, the concentration in recent discussions and research appears to be focused on the following fields: **AI and Machine Learning Focus**:

TABLE II
EXTRACTED KEYWORDS

2018		2019		2020		2021	
word	**TF-IDF**	**word**	**TF-IDF**	**word**	**TF-IDF**	**word**	**TF-IDF**
end	0.213	paraphrasing	0.200	pruning	0.054	learning	0.034
adaptive	0.129	nuclear	0.115	structured	0.054	training	0.028
keyboard	0.129	clinical	0.082	training	0.048	transformer	0.024
system	0.129	concept	0.082	baltic	0.046	tuning	0.024
data	0.119	contextual	0.082	modeling	0.045	semantic	0.023
recognition	0.118	embeddings	0.082	generation	0.045	knowledge	0.022
speech	0.118	enhancing	0.082	story	0.045	understanding	0.021
machine	0.109	extraction	0.082	conversational	0.044	pre	0.020
monolingual	0.109	adjectives	0.058	play	0.044	trained	0.020
neural	0.109	captains	0.058	plug	0.044	fine	0.020
systematic	0.109	challenge	0.058	knowledge	0.043	bert	0.019
translation	0.109	commonsense	0.058	alternative	0.038	generation	0.018
russian	0.107	dataset	0.058	entity	0.038	data	0.017
cast	0.079	objects	0.058	evaluation	0.038	text	0.017
containing	0.079	reasoning	0.058	holms	0.038	shot	0.017
explosives	0.079	submarines	0.058	latvian	0.038	classification	0.017
heat	0.079	application	0.055	named	0.038	multilingual	0.016
ignition	0.079	common	0.055	recognition	0.038	programming	0.016
loading	0.079	driven	0.055	summary	0.038	controlling	0.015
microscopic	0.079	eeg	0.055	answering	0.036	dialogue	0.015

2022		2023		2024		2017-2024	
word	**TF-IDF**	**word**	**TF-IDF**	**word**	**TF-IDF**	**word**	**TF-IDF**
ai	0.027	chatgpt	0.019	chatgpt	0.020	chatgpt	0.019
code	0.024	ai	0.016	ai	0.018	ai	0.016
learning	0.022	learning	0.015	learning	0.014	learning	0.014
machine	0.022	generation	0.012	intelligence	0.014	generative	0.012
artificial	0.021	generative	0.012	artificial	0.014	intelligence	0.012
intelligence	0.021	text	0.011	generative	0.014	generation	0.012
programming	0.020	knowledge	0.011	analysis	0.012	artificial	0.012
gpt	0.020	intelligence	0.011	generation	0.011	text	0.011
search	0.020	artificial	0.011	data	0.010	knowledge	0.011
human	0.019	gpt	0.010	text	0.010	data	0.010
writing	0.018	data	0.010	evaluation	0.009	gpt	0.010
generation	0.018	analysis	0.008	knowledge	0.009	analysis	0.009
shot	0.017	evaluation	0.008	education	0.009	evaluation	0.008
conference	0.017	human	0.008	gpt	0.009	medical	0.008
proceedings	0.017	education	0.008	medical	0.009	education	0.008
natural	0.017	medical	0.008	prompt	0.008	human	0.008
translation	0.017	detection	0.007	performance	0.008	prompt	0.007
system	0.017	natural	0.007	human	0.008	code	0.007
data	0.017	prompt	0.007	information	0.007	detection	0.007
prompts	0.016	exploring	0.007	code	0.007	exploring	0.007

The presence of terms such as *AI*, *machine learning*, *deep learning*, *neural*, *training*, *prediction*, and *dataset* suggests a strong emphasis on the development and application of AI technologies. This is indicative of ongoing advancements in AI capabilities and their broad applications across various domains. **GPT and Generative Models**: the frequent mention of *GPT*, *generative*, and *transformer* indicates specific interest in transformer-based models, especially those related to OpenAI's GPT series. These models are central to discussions on natural language processing, showcasing their significant role in advancing conversational AI and text generation. Moreover, it is evident that, albeit many models (e.g., llama, Gemini, Mixtral) have emerged in the last few years, these models fail to gain enough significance to appear among the most frequent keywords, and GPT models remains by far the most widely used model. **Application Areas**: words like *healthcare*, *medical*, *education*, and *legal* suggest that AI's impact is being explored across critical sectors. This aligns with the trend of deploying AI for specialized applications, reflecting a maturation in the field where AI moves from theoretical exploration to practical, impactful applications. **Evaluation and Performance**: terms like *evaluation*, *performance*, *benchmark*, *accuracy*, and *efficiency* highlight the importance of metrics and testing in AI development. This indicates a focus on not just creating AI models but also on validating their effectiveness and efficiency in various settings. **Ethics and Challenges**: the inclusion of terms like *bias*, *ethical*, *challenges*, and *limitations* points to a growing awareness and discussion around the ethical implications of AI. It shows that alongside technological advancements, there is a parallel conversation about the societal impacts of AI, including concerns about bias, fairness, and the transparency of AI systems. **Technical Aspects and Innovation**: words like *architecture*, *code*, *framework*, and *engineering* imply a robust discussion around the technical underpinnings and innovations in AI.

This might involve exploring new model architectures, enhancing existing frameworks, and improving the integration of AI into software engineering practices. **Knowledge and Information**: terms such as *knowledge*, *information*, *data*, *semantic*, and *extraction* underscore the centrality of data and knowledge management in AI. This reflects the AI field's reliance on vast amounts of data to train models and the ongoing efforts to refine information extraction and semantic analysis techniques.

This latter deeper analysis on frequent and significant terms depict a domain that is growing not only technologically but also in its complex interactions with important ethical, practical, and societal issues. The emphasis covers everything from basic technologies and methods to applications that affect daily life and essential sectors.

In conclusion, this review showcased the broad application of LLMs across various sectors, highlighting their potential to enhance efficiency, accuracy, and productivity. Notable strengths observed include their ability to generate coherent and relevant texts across diverse contexts and to learn effectively from minimal examples. Nonetheless, limitations such as the potential for generating inaccurate or incoherent outputs, perpetuating biases from training data, and lacking a true understanding of human language and context were also noted. This comprehensive analysis provided an overview of the subclassifications, fields, strengths, and weaknesses of LLM use, indicating their significant yet recent impact in transforming industry paradigms and replacing less efficient systems or improving them thanks to their integration.

Acknowledgment

The research leading to these results has received partial funding from the Italian Ministry of Education and Research (MIUR) in the framework of the ForeLab project (Departments of Excellence), from PNRR - M4C2 - Investimento 1.3, Partenariato Esteso PE00000013- "FAIR-Future Artificial Intelligence Research" - Spoke 1 "Human-centered AI", funded by the European Commission under the NextGeneration EU programme, and from the PRIN grant no. 2022ALBSWX of the Italian Ministry of University and Research.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[2] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, *et al.*, "Improving language understanding by generative pre-training," 2018.

[3] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," 2020.

[4] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J.-Y. Nie, and J.-R. Wen, "A survey of large language models," 2023.

[5] S. Minaee, T. Mikolov, N. Nikzad, M. Chenaghlu, R. Socher, X. Amatriain, and J. Gao, "Large language models: A survey," 2024.

[6] H. Naveed, A. U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Akhtar, N. Barnes, and A. Mian, "A comprehensive overview of large language models," 2024.

[7] J. K. Kim, M. Chua, M. Rickard, and A. Lorenzo, "Chatgpt and large language model (llm) chatbots: The current state of acceptability and a proposal for guidelines on utilization in academic medicine," *Journal of Pediatric Urology*, vol. 19, no. 5, pp. 598–604, 2023.

[8] D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu, *et al.*, "Palm-e: An embodied multimodal language model," *arXiv preprint arXiv:2303.03378*, 2023.

[9] A. Bucker, L. Figueredo, S. Haddadinl, A. Kapoor, S. Ma, and R. Bonatti, "Reshaping robot trajectories using natural language commands: A study of multi-modal data alignment using transformers," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 978–984, IEEE, 2022.

[10] M. Benary, X. D. Wang, M. Schmidt, D. Soll, G. Hilfenhaus, M. Nassir, C. Sigler, M. Knödler, U. Keller, D. Beule, *et al.*, "Leveraging large language models for decision support in personalized oncology," *JAMA Network Open*, vol. 6, no. 11, pp. e2343689–e2343689, 2023.

[11] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting, "Large language models in medicine," *Nature Medicine*, vol. 29, no. 8, p. 1930 – 1940, 2023. Cited by: 134.

[12] T. Yamazaki, K. Yoshikawa, T. Kawamoto, T. Mizumoto, M. Ohagi, and T. Sato, "Building a hospitable and reliable dialogue system for android robots: a scenario-based approach with large language models," *Advanced Robotics*, vol. 37, no. 21, pp. 1364–1381, 2023.

[13] N. Yoshikawa, M. Skreta, K. Darvish, S. Arellano-Rubach, Z. Ji, L. Bjørn Kristensen, A. Z. Li, Y. Zhao, H. Xu, A. Kuramshin, *et al.*, "Large language models for chemistry robotics," *Autonomous Robots*, vol. 47, no. 8, pp. 1057–1086, 2023.

[14] T. Onorati, Á. Castro-González, J. C. del Valle, P. Díaz, and J. C. Castillo, "Creating personalized verbal human-robot interactions using llm with the robot mini," in *International Conference on Ubiquitous Computing and Ambient Intelligence*, pp. 148–159, Springer, 2023.

[15] G. J. Serfaty, V. O. Barnard, and J. P. Salisbury, "Generative facial expressions and eye gaze behavior from prompts for multi-human-robot interaction," in *Adjunct Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–3, 2023.

[16] D. Umerenkov, G. Zubkova, and A. Nesterov, "Deciphering diagnoses: How large language models explanations influence clinical decision making," 2023.

[17] N. Mehandru, B. Y. Miao, E. R. Almaraz, M. Sushil, A. J. Butte, and A. Alaa, "Evaluating large language models as agents in the clinic," *npj Digital Medicine*, vol. 7, p. 84, Apr 2024.

[18] M. Hosseini, C. A. Gao, D. M. Liebovitz, A. M. Carvalho, F. S. Ahmad, Y. Luo, N. MacDonald, K. L. Holmes, and A. Kho, "An exploratory survey about using chatgpt in education, healthcare, and research," *Plos one*, vol. 18, no. 10, p. e0292216, 2023.

[19] E. Kasneci, K. Sessler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier, S. Krusche, G. Kutyniok, T. Michaeli, C. Nerdel, J. Pfeffer, O. Poquet, M. Sailer, A. Schmidt, T. Seidel, M. Stadler, J. Weller, J. Kuhn, and G. Kasneci, "Chatgpt for good? on opportunities and challenges of large language models for education," *Learning and Individual Differences*, vol. 103, p. 102274, 2023.

[20] S. Wu, O. Irsoy, S. Lu, V. Dabravolski, M. Dredze, S. Gehrmann, P. Kambadur, D. Rosenberg, and G. Mann, "Bloomberggpt: A large language model for finance," 2023.

[21] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. de las Casas, E. B. Hanna, F. Bressand, G. Lengyel, G. Bour, G. Lample, L. R. Lavaud, L. Saulnier, M.-A. Lachaux, P. Stock, S. Subramanian, S. Yang, S. Antoniak, T. L. Scao, T. Gervet, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed, "Mixtral of experts," 2024.

[22] J. Demšar, T. Curk, A. Erjavec, Črt Gorup, T. Hočevar, M. Milutinovič, M. Možina, M. Polajnar, M. Toplak, A. Starič, M. Štajdohar, L. Umek, L. Žagar, J. Žbontar, M. Žitnik, and B. Zupan, "Orange: Data mining toolbox in python," *Journal of Machine Learning Research*, vol. 14, pp. 2349–2353, 2013.

[23] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information Processing & Management*, vol. 24, no. 5, pp. 513–523, 1988.