Unraveling the landscape of large language models: a systematic review and future perspectives

Qinxu Ding, Ding Ding and Yue Wang
School of Business, Singapore University of Social Sciences, Singapore, Singapore
Chong Guan
Centre for Continuing and Professional Education, Singapore University of Social Sciences, Singapore, Singapore, and
Bosheng Ding
Nanyang Technological University, Singapore, Singapore

Received 5 August 2023
Revised 4 September 2023
Accepted 5 September 2023

Abstract

Purpose – The rapid rise of large language models (LLMs) has propelled them to the forefront of applications in natural language processing (NLP). This paper aims to present a comprehensive examination of the research landscape in LLMs, providing an overview of the prevailing themes and topics within this dynamic domain.
Design/methodology/approach – Drawing from an extensive corpus of 198 records published between 1996 to 2023 from the relevant academic database encompassing journal articles, books, book chapters, conference papers and selected working papers, this study delves deep into the multifaceted world of LLM research. In this study, the authors employed the BERTopic algorithm, a recent advancement in topic modeling, to conduct a comprehensive analysis of the data after it had been meticulously cleaned and preprocessed. BERTopic leverages the power of transformer-based language models like bidirectional encoder representations from transformers (BERT) to generate more meaningful and coherent topics. This approach facilitates the identification of hidden patterns within the data, enabling authors to uncover valuable insights that might otherwise have remained obscure. The analysis revealed four distinct clusters of topics in LLM research: "language and NLP", "education and teaching", "clinical and medical applications" and "speech and recognition techniques". Each cluster embodies a unique aspect of LLM application and showcases the breadth of possibilities that LLM technology has to offer. In addition to presenting the research findings, this paper identifies key challenges and opportunities in the realm of LLMs. It underscores the necessity for further investigation in specific areas, including the paramount importance of addressing potential biases, transparency and explainability, data privacy and security, and responsible deployment of LLM technology.
Findings – The analysis revealed four distinct clusters of topics in LLM research: "language and NLP", "education and teaching", "clinical and medical applications" and "speech and recognition techniques". Each cluster embodies a unique aspect of LLM application and showcases the breadth of possibilities that LLM technology has to offer. In addition to presenting the research findings, this paper identifies key challenges and opportunities in the realm of LLMs. It underscores the necessity for further investigation in specific areas, including the paramount importance of addressing potential biases, transparency and explainability, data privacy and security, and responsible deployment of LLM technology.
Practical implications – This classification offers practical guidance for researchers, developers, educators, and policymakers to focus efforts and resources. The study underscores the importance of addressing challenges in LLMs, including potential biases, transparency, data privacy, and responsible deployment. Policymakers can utilize this information to shape regulations, while developers can tailor technology development based on the diverse applications identified. The findings also emphasize the need for interdisciplinary collaboration and highlight ethical considerations, providing a roadmap for navigating the complex landscape of LLM research and applications.

Journal of Electronic Business & Digital Economics
Vol. 3 No. 1, 2024
pp. 3-19
Emerald Publishing Limited
e-ISSN: 2754-4222
p-ISSN: 2754-4214
DOI 10.1108/JEBDE-08-2023-0015

Originality/value – This study stands out as the first to examine the evolution of LLMs across such a long time frame and across such diversified disciplines. It provides a unique perspective on the key areas of LLM research, highlighting the breadth and depth of LLM's evolution.

Keywords Large language models, Natural language processing, Artificial intelligence, Topic modeling, BERTopic, ChatGPT

Paper type Research paper

1. Introduction

The advent of large language models (LLMs), including conversational generation tools like OpenAI's ChatGPT and Google's Bard, has revolutionized the field of natural language processing (NLP). These models not only offer highly informative and integrated conversations to users, but also have the capability to develop code, conduct code reviews and fix bugs, among other functionalities (Hsu & Ching, 2023). LLMs have also been applied to various domains and tasks, such as education, healthcare, speech recognition, information retrieval, text summarization and dialog systems. LLMs have shown impressive performance and potential in generating and processing natural language texts for diverse purposes and audiences. However, LLMs also pose significant challenges and risks, such as data quality, model reliability, ethical issues and social impacts. Therefore, it is important to understand the current state and future directions of LLM research and application. This paper presents a comprehensive analysis of the predominant themes and topics addressed in previous research concerning LLMs, drawing from a rich corpus of 198 records published between 1996 to 2023.

The records, which include journal articles, books, book chapters, conference papers and selected working papers, were gathered from a wide range of academic databases, such as EBSCO, Cambridge Journals, Elsevier, Emerald, Institute of Electrical and Electronics Engineers (IEEE), Journal Storage (JSTOR), Nature, Social Science Research Network (SAGE), Springer, Taylor & Francis and Wiley. The analysis was conducted using the BERTopic algorithm, a topic modeling technique that enabled us to identify four distinct clusters within LLM research: "Language and Natural Language Processing", "Education and Teaching", "Clinical and Medical Applications" and "Speech and Recognition Techniques".

Each cluster represents a unique aspect of LLM application, demonstrating the wide-ranging potential of LLM technology. The "Language and Natural Language Processing" cluster delves into the technical aspects of LLMs, exploring key findings, techniques and challenges. The "Education and Teaching" cluster examines the transformative potential of LLMs in automating and enhancing educational tasks, while also addressing associated ethical and practical challenges. The "Clinical and Medical Applications" cluster investigates the promising potential of LLMs in healthcare, with a focus on their use in medical education and patient care. Lastly, the "Speech and Recognition Techniques" cluster explores the use of LLMs in speech recognition and other related applications.

This study stands out as the first to examine the evolution of LLMs across such a long time frame and such diversified disciplines. It provides a unique perspective on the key areas of LLM research and application over the years, highlighting the breadth and depth of LLM's evolution.

In addition to presenting the research findings, this paper also identifies key challenges and opportunities in the realm of LLMs. It underscores the necessity for further investigation in specific areas, including explainability, robustness, cross-modal and multi-modal generation and interactive co-creation. Moreover, the paper highlights the paramount importance of addressing data privacy, security and responsible deployment of LLM technology.

In the following sections, this paper will detail the literature review, and explain the research methodology, data source and analysis techniques used in this study on LLMs.

It will also present and discuss the results in relation to existing theories and practices. Finally, it will suggest some future research directions based on the findings and implications of the study. The aim is to provide a clear and detailed understanding of the research process, outcomes and their significance.

2. Literature review

Some review papers take a holistic view of the field of LLM, from different perspectives and time frames. An up-to-date and comprehensive review of the literature on LLMs from a technical and engineering perspective was offered (Zhao *et al.*, 2023), which can be a useful resource for both researchers and engineers who are interested in ChatGPT or other LLM-based applications. They reviewed the background, key findings and main techniques for LLMs, such as scaling laws, emergent abilities and alignment tuning. They also covered four major aspects of LLMs, namely pretraining, adaptation tuning, utilization and capacity evaluation, and summarized the recent progress and challenges in each aspect. Moreover, they discussed the practical guide for prompt design, which is the key interface for accessing and using LLMs, and the applications of LLMs in several representative domains, such as dialog systems, information retrieval, code generation and education.

Fan *et al.* (2023) conducted a bibliometric review of LLMs research from 2017 to 2023. They analyzed the trends, topics, challenges and applications of LLMs in natural language processing (NLP) and related fields using various bibliometric methods and tools. They collected and visualized the data from 1,672 publications from different countries, institutions, journals and conferences. They found that LLM research has grown rapidly in the past seven years, especially in 2020 and 2021, with more than 1,000 publications in these two years alone. They also found that LLM research is mainly driven by the advances in transformer-based architectures, such as bidirectional encoder representations from transformers (BERT) and Generative Pre-trained Transformer (GPT), and their variants and extensions. LLMs are applied to a wide range of NLP tasks, such as text generation, question answering, natural language understanding, machine translation, sentiment analysis, text summarization and dialog systems. Besides, LLM research is highly collaborative and interdisciplinary, with many co-authorship networks and cross-domain applications.

Other review papers look at a subgroup of the LLM literature, and focus on the advancement of certain selected topics, like education, healthcare research, etc. For example, Yan *et al.* (2023) looked at how LLM was used in education. They conducted a systematic scoping review of 118 peer-reviewed papers published since 2017 to pinpoint the current state of research on using LLMs to automate and support educational tasks. They identified 53 use cases for LLMs in education, categorized into nine main categories: profiling/labeling, detection, grading, teaching support, prediction, knowledge representation, feedback, content generation and recommendation. They also discussed several practical and ethical challenges, such as low technological readiness, lack of replicability and transparency, and insufficient privacy and beneficence considerations.

Sallam (2023) is a systematic review of the role and limitations of ChatGPT. The paper identifies the potential benefits of ChatGPT, such as improving scientific writing, analyzing large datasets, assisting radiologists and providing personalized learning and medicine. The paper also discusses the possible risks and concerns of ChatGPT, such as ethical issues, factual inaccuracies, plagiarism, transparency problems, legal issues and infodemic risk. The paper emphasizes the need for careful and responsible use of ChatGPT, as well as the importance of open data and open science publishing to ensure the safety and quality of health care and research. On the other hand, Chang *et al.* (2023) conducted a comprehensive survey on the evaluation of LLMs that can generate and process natural language texts for various tasks and domains. The paper reviews the existing evaluation methods and metrics

for LLMs from three dimensions: what to evaluate, where to evaluate and how to evaluate. The paper covers a wide range of evaluation tasks, such as natural language processing, reasoning, robustness, ethics, biases, trustworthiness, social science, natural science and engineering, medical applications, agent applications and other applications. The paper also summarizes the existing evaluation datasets and benchmarks for LLMs, and discusses their advantages and limitations.

Wei *et al.* (2022a) took a different perspective and discussed the phenomenon of emergent abilities of LLMs, which are abilities that are not present in smaller models but are present in larger models. The paper surveys several examples of emergent abilities from prior work, such as few-shot prompting, instruction following, multi-step reasoning and model calibration. The paper uses various sources of data, such as scaling curves, benchmarks, datasets and metrics, to identify and evaluate emergent abilities. The paper also uses a systematic definition of emergence and a focused scope of LLMs to provide a clear and consistent framework for discussing emergent abilities.

The literature reviews on LLMs have some limitations that may affect their validity and applicability. One limitation is the range of time that they cover, which is mostly limited to recent years, especially 2020 and 2021. This may exclude some earlier works that are still relevant and influential in the field of LLMs. Another limitation is the emphasis that they put on certain topics or aspects of LLMs, which may not reflect the diversity and complexity of the field. For example, some reviews may focus more on the technical or engineering aspects of LLMs, while others may focus more on the ethical or social aspects of LLMs. This may result in a partial or incomplete picture of the state-of-the-art and the challenges of LLMs.

Our paper addresses these limitations by conducting a systematic and comprehensive review of the literature on LLMs from a multidisciplinary perspective. We cover a wide range of time frames, from 1996 to 2023, and include both peer-reviewed and preprint publications from different sources and domains. We use rigorous and transparent methods to select, analyze and synthesize the records, and we provide a critical appraisal and a meta-analysis of the evidence. We also address the ethical, social and legal implications of LLMs in depth and provide clear and actionable recommendations for their responsible use and regulation.

LLMs have evolved significantly in terms of their development and architecture, thanks to the breakthrough of the transformer architecture by (Vaswani *et al.*, 2017). This architecture enabled the creation of the GPT series models, from GPT-1 to GPT-4 (OpenAI, 2023; Radford *et al.*, 2019), which demonstrated remarkable performance in natural language processing. Among them, InstructGPT showed superior results compared to the larger GPT-3 model, despite having fewer parameters. Another notable contribution in this field is the "Constitutional AI" method proposed by Claude from Anthropic (Bai *et al.*, 2022). Moreover, to address the computational cost issues, Meta's LLaMA model presented an optimized approach for various inference budgets.

LLMs have a wide range of practical applications in different domains. In education, LLMs are considered as potential game-changers. Bonner, Lege, and Frazier (2023) highlighted their potential to revolutionize educational experiences. Chaudhry, Cukurova, and Luckin (2022) stressed the importance of transparency in Artificial Intelligence (AI) applications, especially in educational settings. Chechitelli (2023) also provided insights into the role of AI in plagiarism detection, while Condor, Litster, and Pardos (2021) explored AI's transformative potential in automating educational assessments. In healthcare, studies such as those by Kung *et al.* (2023) and Liévin, Hother, and Winther (2022) illustrated the promising capabilities of ChatGPT in medical education and reasoning. Translation and speech recognition have also benefited from LLMs, with significant advancements in machine translation (Vaswani, Zhao, Fossum, & Chiang, 2013) and speech recognition (Kim *et al.*, 2020).

However, LLMs also pose some challenges and ethical issues that need to be addressed. A paramount concern in the LLM community is model safety. Research on GPT-4 has

explored safety-relevant reinforcement learning with human feedback (RLHF) and rule-based reward models (RBRMs) to address these concerns. However, challenges persist, such as the "hallucination" phenomenon observed in ChatGPT, where the model can produce misleading or nonsensical answers. Additionally, the potential for LLMs like ChatGPT to reproduce biases from their training data remains a significant ethical concern.

Efficiency in LLMs is not only about computational prowess but also about the effective utilization of prompts and training strategies. The significance of prompt setting in LLMs has been highlighted in various studies, with chain-of-thought prompting emerging as a promising technique (Wei *et al.*, 2022a, b). Mayer, Ludwig, and Brandt (2023) have also demonstrated the efficacy of prompt-based learning for domain-specific tasks. To address the storage and computational challenges of LLMs, Schwenk, Rousseau, and Attik (2012) advocated for efficient training strategies, emphasizing the potential benefits of modern multi-core computers. In this study, we employ a unique approach by leveraging a topical model to examine the primary focus areas of papers on LLMs over a considerable time span. Our analysis reveals four primary categories which capture the majority of these topics: "language and NLP", "education and teaching", "clinical and medical applications" and "speech and recognition techniques". We provide a detailed summary of the core findings from representative papers within each category, underscoring that the first category analyzes LLMs from a technical perspective, while the rest categories focus on their practical applications, suggesting potential interconnections.

3. Methodology

This paper presents the first comprehensive investigation into the prevalent themes and subjects explored in previous research related to LLMs, utilizing topic modeling techniques. The dataset encompasses records sourced from reputable academic databases such as EBSCO, Cambridge Journals, Elsevier, Emerald, IEEE, JSTOR, Nature, SAGE, Springer, Taylor & Francis and Wiley. These records comprise abstracts and citations from peer-reviewed literature across diverse domains, spanning the period from 1985 to 2023. The dataset includes a variety of scholarly works, such as journal articles, books, book chapters, conference papers and selected working papers, specifically targeting LLM-related topics. The search was conducted using two sets of keywords: "LLM" and "Large" + "Language" + "Model," with the latter yielding a subset within the broader list obtained from the former. An observable surge in publications occurred post-2021, constituting more than 90% of the dataset. Subsequently, a thorough examination of the abstracts within the collected records was performed, leading to the exclusion of unqualified entries. Ultimately, 198 records were retained for the final analysis.

Topic modeling stands as a widely utilized method in the domains of NLP. Its primary objective involves extracting latent topics from a given corpus, thereby aiding in the comprehension of underlying themes and structural patterns within the textual data. Consequently, this technique facilitates effective organization and analysis of the information, contributing to enhanced knowledge discovery and insights.

While traditional topic modeling algorithms like latent Dirichlet allocation (LDA) (Blei, Ng, & Jordan, 2003) have been widely applied in various contexts, there are several limitations that can limit the efficacy, such as issues related to handling noisy data, topic overlapping and limited interpretability. To address these limitations, BERTopic (Grootendorst, 2022) has emerged as a promising alternative. BERTopic is a sophisticated topic modeling algorithm built upon the BERT architecture that was developed recently (Devlin, Chang, Lee, & Toutanova, 2019). Notably, BERTopic is well-regarded for its ability to generate coherent topics and exhibit competitive performance across diverse benchmark evaluations. It surpasses classical models like LDA and non-negative matrix factorization (NMF) (Févotte & Idier, 2011) as well as more recent clustering-based approaches to topic modeling, such as

Correlated Topic Model (CTM) (Bianchi, Terragni, & Hovy, 2021) and Top2Vec (Angelov, 2020). Prior studies have compellingly showcased the effectiveness of BERTopic in identifying dimensions and prominent keywords expressed in online reviews (Atzeni, Bacciu, Mazzei, & Prencipe, 2022; Raju *et al.*, 2022) and uncovering research patterns in the field of literature analysis (Fan *et al.*, 2023), highlighting its value and utility in diverse research contexts. Therefore, the abstracts of the selected papers in this study were analyzed using BERTopic to generate a list of interpretable topics from the previous literature.

In particular, BERTopic employs a three-step approach to create topic representations. Firstly, it converts each document into an embedding representation using a pretrained language model. Next, to improve the clustering process, it reduces the dimensionality of the resulting embeddings before clustering them. Finally, a customized class-based variation of Term Frequency-Inverse Document Frequency (TF-IDF) is utilized to extract topic representations from the document clusters (Bafna *et al.*, 2016). These three distinct steps facilitate a flexible topic model suitable for various applications, including dynamic topic modeling.

Prior to topic modeling, the textual data extracted from the abstracts underwent text preprocessing to enhance the quality of the generated topic models. This stage involves several key steps, including: a) Lowercasing: Transforming all text to lowercase to achieve uniformity and eliminate variations arising from capitalization; b) Tokenization: Segmenting the text into individual words or tokens to facilitate further analysis; c) Lemmatization: Reducing words to their base or dictionary form, for instance, converting "running" to "run"; d) Stitching Tokens: Reassembling the lemmatized tokens to form cleaned text, ready for subsequent analysis.

1. Document embeddings

During this phase, BERT-based embeddings are produced for individual documents, specifically abstracts, employing the BERT model. These embeddings effectively encapsulate contextual nuances and semantic nuances inherent in the textual content of the abstracts. To accomplish this, BERTopic adopts the utilization of sentence transformers, a pertinent tool. Within this context, we opted to employ the default embedding model offered by BERTopic, denoted as all-MiniLM-L6-v2 as outlined in the pretrained models section of the sentence-transformers documentation. This particular model facilitates the transformation of sentences and paragraphs into a 384-dimensional dense vector space, a representation conducive to tasks such as clustering or semantic exploration.

2. Document clustering

Upon acquiring the embeddings for the abstracts of the obtained documents, the application of clustering algorithms assumes a pivotal role, aiming to categorize the documents into coherent clusters, each encapsulating a distinct thematic focus. However, the challenge arises when dealing with datasets characterized by high-dimensional attributes, as an augmentation in dimensionality has been observed to lead the proximity to the nearest data point to approximate the distance to the farthest one. As a result, the conventional notion of spatial proximity becomes elusive within high-dimensional spaces, resulting in a convergence of diverse distance metrics. To mitigate this concern, various clustering methodologies have been devised to counteract the repercussions of the "curse of dimensionality." Nevertheless, a more direct strategy involves the reduction of embedding dimensionality.

Although principal component analysis (PCA) (Jolliffe & Cadima, 2016) and t-distributed stochastic neighbor embedding (t-SNE) (Maaten & Hinton, 2008) stand as widely acknowledged techniques for dimensionality reduction, an alternative approach termed uniform manifold approximation and projection (UMAP) (McInnes, Healy, & Melville, 2020) has surfaced, demonstrating superior retention of both local and global characteristics

intrinsic to high-dimensional data during the projection onto lower dimensions. Notably, UMAP presents the advantage of versatility across varied dimensional spaces of language models, unburdened by computational limitations imposed on embedding dimensions. Consequently, we opt to leverage the capabilities of UMAP to effectuate the reduction in dimensionality of the document embeddings engendered within this stage.

When clustering the thusly reduced embeddings, BERTopic deploys a modified rendition of the hierarchical density-based spatial clustering of applications with noise (HDBSCAN) (McInnes, Healy, & Astels, 2017). This approach extends the foundational underpinnings of Density-Based Spatial Clustering of Applications with Noise (DBSCAN) (Ester *et al.*, 1996) by discerning clusters characterized by differing densities through the employment of a hierarchical clustering algorithm. By employing a soft-methodology, HDBSCAN models cluster in a manner conducive to the incorporation of noise modeling, effectively treating noise as outliers. Consequently, this approach forestalls the erroneous allocation of unrelated documents to specific clusters, ultimately enhancing the representation of topics.

3. Topic Representation

The final stage pertains to the derivation of topic depictions of the abstracts of the obtained documents. For every cluster formulated in the preceding phase, BERTopic calculates the Class-based TF-IDF (c-TF-IDF) scores for each term within the cluster. Essentially, terms characterized by the highest c-TF-IDF scores are considered the most representative of their respective topics. Here, we used the built-in reduce_frequent_words function of BERTopic to reduce frequent words in the c-TF-IDF representation. Essentially, this approach entails a reduction in the weightage accorded to words that appear very frequently across the corpus of documents.

For a term **x** within class **c**:

$$w_{x,c} = \left\| tf_{x,c} \right\| \times log\left(1 + \frac{A}{f_x}\right)$$

where.

$tf_{x,c} = frequency\ of\ word\ \boldsymbol{x}\ in\ class\ \boldsymbol{c}$

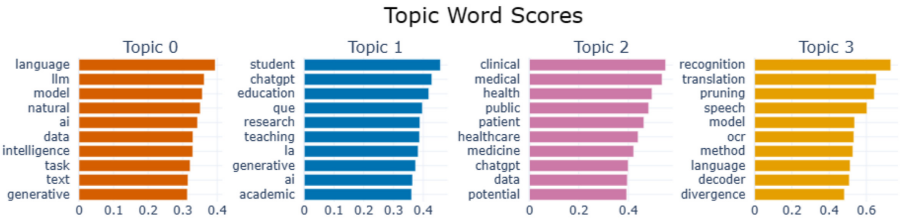
$f_x = frequency\ of\ word\ \boldsymbol{x}\ across\ all\ classes$

$A = average\ number\ of\ words\ per\ class$

4. Result

4.1 Determining the number of topics

Following the implementation and training of the BERTopic model on the dataset, we have effectively derived four distinct topics. As depicted in Figure 1, the visualization illustrates



Source(s): Figure by authors

Figure 1. The illustration of the topic words for the *four* topics derived by BERTopic

that BERTopic efficiently extracts a set of lexemes representative of each topic. These lexical items, commonly referred to as "topic words", assume a vital role in encapsulating the principal themes or concepts inherent within the corpus pertaining to a specific topic. BERTopic employs a scoring mechanism to assess these topic words, a metric that conveys the relative significance of individual words with respect to a given topic. The computation of these scores is underpinned by sophisticated algorithms that consider the contextual usage and distribution patterns of words within the corpus. Words that exhibit close proximity and contribute substantially to the semantic essence of a particular topic tend to garner higher scores, signifying their relevance. Conversely, words deemed less germane or making only marginal contributions to the semantic delineation of the topic receive lower scores.

An intertropic distance map, visually depicted in Figure 2, serves to delineate the interconnections among topics within a two-dimensional expanse. The intertopic distance map offers an understanding of the relationships between various topics, based on their semantic resemblances. It visualizes the topics in a two-dimensional plane, wherein the separation between topics represents their semantic likeness. Topics situated in proximity exhibit higher similarity concerning the words and context they encompass, whereas topics situated at a greater distance are more dissimilar. The resultant map provides insights into the proximate associations and hierarchical structure of topics, thus augmenting the depth of comprehension regarding the topic arrangement intrinsic to the dataset.

Upon conducting a comprehensive evaluation incorporating the above, we deduced four distinct clusters of topics, which are tabulated in Table 1.

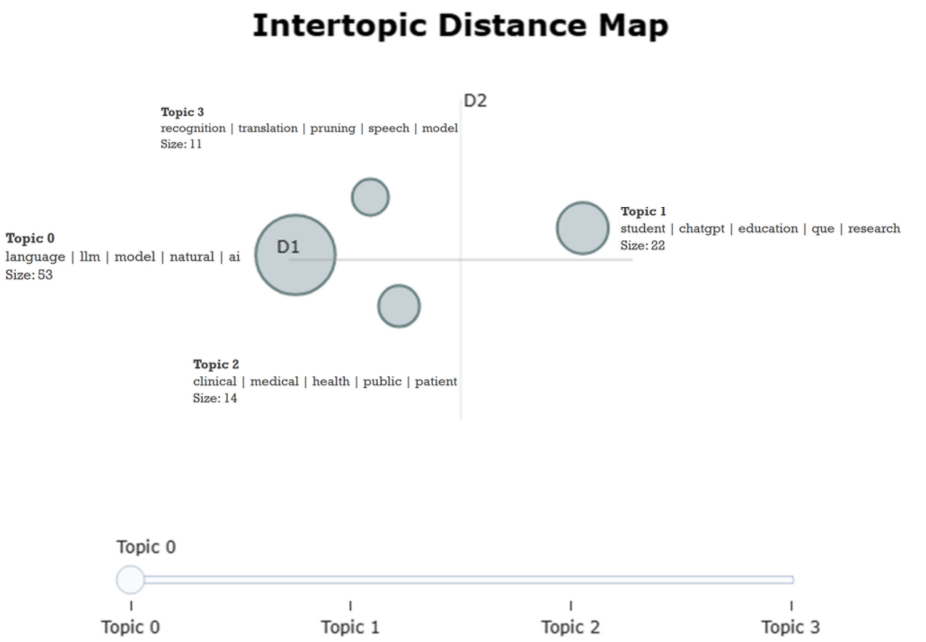


Figure 2.
Intertopic distance map for the 4 topics derived by BERTopic

Source(s): Figure by authors

5. Discussion

5.1 Language and Natural Language Processing (NLP)

Enhance the performance of LLMs on NLP tasks: One of the leading research aspects of LLMs is to improve their performance on NLP tasks by increasing their scale. We can clearly see this trend by observing the evolutionary path of the GPT series models. These models, such as GPT-1, GPT-2, GPT-3, InstructGPT and GPT-4 (Brown *et al.*, 2020; OpenAI, 2023; Ouyang *et al.*, 2022; Radford, Narasimhan, Salimans, & Sutskever, 2018; Radford *et al.*, 2019), are based on the architecture of the transformer (Vaswani *et al.*, 2017), which is one of the famous architectures for developing LLMs and is well-known for its attention mechanisms, resulting in a simpler architecture than the recurrent neural networks (RNNs) (Schuster & Paliwal, 1997). GPT-1 model (Radford *et al.*, 2018) was proposed to build a generative pre-training language model for various natural language tasks such as question answering, document classification, etc. It has 117 million parameters. The motivation of this work is to build the model with a diverse corpus of unlabeled text, as there is more unlabeled text than labeled text in the real world. A relatively small set of labeled text would be used to fine-tune the generative pretraining model for specific tasks. Although GPT-1 model performs well on these tasks, it still relies on a labeled dataset for each specific task to finetune the model. The GPT-2 model (Radford *et al.*, 2019) was proposed to further reduce the need for labeled datasets. It is a 1.5 billion transformer and uses a zero-shot setting, demonstrating that LLMs can learn various natural language tasks without explicit supervision. Compared with GPT-1 and GPT-2 models, GPT-3 model (Brown *et al.*, 2020) has 175 billion parameters, which is much larger and improves the performance of LLMs further. GPT-4 model (OpenAI, 2023), a large-scale and multimodal model, has 1.8 trillion parameters, which is over 10 times larger than GPT-3. It performs better than GPT-3 model on a diverse set of benchmarks, such as exams for humans.

Align LLMs' output with human instructions: Although increasing the scale can make LLMs perform well on NLP tasks, it still cannot ensure that the generated outputs align with humans' instructions, resulting in unsatisfied user interactions and incorrect information. To mitigate this problem, InstructGPT (Ouyang *et al.*, 2022) was proposed, which uses supervised learning to fine-tune the GPT-3 model with human prompts. Meanwhile, reinforcement learning is used to reward the LLM if its results are aligned with human prompts. In fact, a 1.3B parameter InstructGPT model is even better than a 175B GPT-3 model by human evaluations. These results further demonstrate that an LLM cannot only rely on its large scale or complex structure.

Safe and harmless LLMs: On the other hand, the safety of LLMs also attracts researchers' attention. Although ChatGPT or InstructGPT has the incredible performance to align with users' prompts, no one can make sure the results are not harmful to humans. For example, if unsafe inputs are given, the model may instruct on committing crimes. To mitigate this problem, GPT-4 model uses two components named safety-relevant RLHF and RBRMs. Meanwhile, Claude (Bai *et al.*, 2022), the LLM developed to beat ChatGPT by anthropic, uses a

Topic	Name	Representation
0	0_language_llm_model_natural	["language", "llm", "model", "natural", "ai", "data", "intelligence", "task", "text", "generative"]
1	1_student_chatgpt_education_que	["student", "chatgpt", "education", "que", "research", "teaching", "la", "generative", "ai", "academic"]
2	2_clinical_medical_health_public	["clinical", "medical", "health", "public", "patient", "healthcare", "medicine", "chatgpt", "data", "potential"]
3	3_recognition_translation_pruning_speech	["recognition", "translation", "pruning", "speech", "model", "ocr", "method", "language", "decoder", "divergence"]

Source(s): Table by authors

Table 1. Summary of the clusters of topics

"constitutional AI" method to improve the safety of LLMs' outputs. Compared with ChatGPT, Claude uses a combination of human feedback and AI feedback in the reinforcement learning stage, while ChatGPT only uses human feedback. In fact, the AI feedback of Claude comes from a harmless AI assistant. Using an AI assistant to supervise AI is a potential direction to improve the safety of LLMs. It can also improve efficiency and reduce cost, as collecting human feedback is difficult. However, there are also concerns. For example, would LLMs' effectiveness be significantly affected if we require too much on LLMs' safety? Moreover, is it reliable to use AI to supervise AI? Deeper research is still needed to find a good trade-off between safety and accuracy of LLMs.

LLMs' development with limited cost budget: Cost is another big issue to consider in developing LLMs. If we can ignore the cost, we can quickly increase the scale of model and dataset. We can even hire more people to label the datasets. However, the budget is not ignorable in the real world. Can we still have a good-performance LLM with a smaller dataset or model? Meta's LLaMA model was developed to achieve the best performances considering various inference budgets. It only uses publicly available datasets, while ChatGPT also uses nonpublicly available datasets. LLaMA-13B can outperform GPT-3 (175B) on some natural language benchmarks. It demonstrates that developing a good LLM with a much smaller model scale and solely public datasets is possible. As the model is released to the public, it will also accelerate the development of LLMs. The generative AI community can work on solving the bias and unsafety issues with these open-source models and datasets.

Prompt setting of LLMs: The prompt setting of LLMs should not be ignored. On the one hand, prompt setting can reduce the requirement of a large, labeled dataset used for training the LLMs. For example, Brown *et al.* (2020) compared zero-shot, one-shot and few-shot settings with GPT-3 model and state-of-art fine-tuned models. They demonstrated that using few-shot settings with GPT-3 model can also perform well on NLP tasks. On the other hand, a suitable prompt setting can also improve the ability of LLMs in complexity reasoning. A standard prompt setting, simple input-out pairs, is insufficient to teach LLMs to perform well in reasoning tasks. The chain-of-thought prompting (Wei *et al.*, 2022b), showing LLMs a series of intermediate reasoning steps, was proposed to improve the reasoning ability of LLMs. Moreover, prompt-based learning can also simply be the people's usage of AI. For example, prompt-based learning approaches with LLMs, without updating the weights of LLMs, can easily be used for domain-specific tasks, such as classifying email responses to a problem-solving task (Mayer *et al.*, 2023).

5.2 Large language model in research and education

The rise of generative artificial intelligence has bestowed upon the academic community unprecedented possibilities for advancing education and research. LLMs, such as ChatGPT, have demonstrated exceptional capabilities in understanding and generating human language, leading to their widespread adoption in various domains, including education and research. The current body of literature on LLMs delves into the transformative impact of LLMs in these domains, with a focus on the enriching experiences for students, novel teaching methodologies, empowered research endeavors and interactive learning enabled by LLMs (Liu *et al.*, 2023).

Enhance student learning experiences: Through their natural language processing and generation abilities, LLMs can process vast amounts of educational content and cater personalized information to individual learners (Sallam, 2023). By tailoring educational materials to suit each student's proficiency and learning pace, LLMs foster a more engaging and interactive learning environment. Additionally, LLMs can facilitate interactive learning sessions wherein students can ask questions and receive real-time responses from the AI, promoting self-directed learning and knowledge acquisition.

One of the fascinating applications of LLMs in education is their integration with AI-powered entities like ChatGPT. This interaction allows students to engage in dialog with the AI, effectively transforming the traditional student-teacher dynamic. As ChatGPT operates as an intelligent conversation partner, students can seek clarifications, discuss concepts and receive guidance outside the classroom setting. This continual interaction not only fosters a deeper understanding of academic topics but also nurtures critical thinking and analytical skills (Kasneci *et al.*, 2023).

Provide innovative methodologies in teaching and assessments: From the educator's perspective, LLMs have ushered in innovative methodologies that challenge traditional instructional paradigms. Leveraging LLMs, educators can develop dynamic lesson plans with interactive elements, thereby captivating students' attention and fostering active participation (MacNeil *et al.*, 2022). Moreover, LLMs can act as virtual teaching assistants, providing instant feedback on assignments and performance assessments (Leinonen *et al.*, 2023). This enables educators to identify students' strengths and weaknesses more efficiently, leading to personalized interventions that enhance learning outcomes.

Improve workflow and efficiency in research: In the domain of research, LLMs have revolutionized data analysis, literature review processes and hypothesis generation. LLMs can analyze extensive volumes of academic literature, expediting the research process and aiding researchers in identifying relevant sources. Furthermore, researchers can employ LLMs to generate insightful summaries of research findings, condensing complex information into easily digestible formats. The use of LLMs in research not only saves time and effort but also empowers scholars to explore new avenues for knowledge discovery.

While the applications of LLMs in education and research are promising, the integration of AI in academic spaces warrants careful ethical deliberation. Privacy concerns, data security and potential biases in generated content are challenges that require meticulous attention (Teubner, Flath, Weinhardt, van der Aalst, & Hinz, 2023). Educators and researchers must strike a balance between utilizing LLMs for enhancing academic practices and safeguarding the welfare of students and researchers. Rigorous oversight and ethical guidelines are imperative to mitigate these challenges and ensure the responsible use of LLMs in academia. Furthermore, it is vital to train LLMs on diverse and inclusive datasets to minimize bias and support fair and equitable learning experiences for all students (Kasneci *et al.*, 2023).

The applications of LLMs in education and research present a compelling landscape of possibilities for the academic community. From enriching student learning experiences and improving educational practices to empowering research endeavors and enabling dynamic interactions with AI entities like ChatGPT, LLMs have the potential to transform the academic sphere. Nonetheless, ethical considerations must underpin these applications, ensuring that AI augmentation in education and research aligns with the best interests of students, researchers and the broader academic community. As LLM technology continues to evolve, its integration into academia will undoubtedly reshape the future of teaching, learning and knowledge generation.

5.3 Large language model in medical and public health

The advent of LLMs has brought forth ground-breaking opportunities for the medical and public health sectors. LLMs, with their natural language processing and generation abilities, possess the capacity to process vast amounts of clinical and health-related data, offering valuable insights and augmenting human decision-making (Ufuk, 2023). From the existing literature, we can observe the manifold applications of LLMs in clinical settings, patient care, medical education and research, and public health initiatives, underscoring their potential to revolutionize the healthcare landscape (Kung *et al.*, 2023).

Improve clinical diagnoses and patient treatment: In the realm of medical and clinical practices, LLMs have the potential to reshape the way clinical professionals diagnose and treat

patients. By analyzing extensive medical literature, electronic health records and research papers, LLMs can assist healthcare providers in making more informed decisions and formulating personalized treatment plans (Liévin *et al.*, 2022). Furthermore, LLMs can analyze patient data and suggest potential interventions, aiding in early detection and management of medical conditions. The integration of LLMs into electronic health record systems holds the promise of streamlining clinical workflows, reducing administrative burden and optimizing healthcare delivery. In the area of patient care, LLMs have the capacity to facilitate improved patient interactions and health literacy. Through their natural language understanding capabilities, LLMs can effectively engage with patients, answering their medical queries and providing essential health information. Moreover, LLM-powered virtual assistants can offer round-the-clock support, empowering patients to take a proactive role in managing their health. The integration of LLMs into telemedicine platforms enables remote patient monitoring and enhances access to healthcare services, particularly in underserved regions (Sallam, 2023).

Enhance medical research and education: In the field of medical research and education, LLMs present invaluable tools for data analysis, literature review and hypothesis generation. By processing extensive medical databases and scientific literature, LLMs can identify patterns and associations that might have otherwise been overlooked. This capacity to explore vast volumes of data expeditiously opens up new avenues for medical research and accelerates the pace of scientific discovery. LLMs can also assist researchers in formulating research questions, facilitating experimental design and interpreting results, thus bolstering the quality and efficiency of medical research endeavors. On the other hand, LLMs have also shown immense potential in transforming medical education. These advanced AI models, like ChatGPT, can act as invaluable virtual teaching assistants, providing students with detailed and up-to-date information on various medical topics (Lee, 2023). LLMs offer personalized learning pathways, adapting their responses to suit individual students' needs and learning patterns. By incorporating LLMs into medical education, students can access a wealth of knowledge, receive interactive feedback and engage in dynamic learning experiences, ultimately fostering a more effective and comprehensive understanding of biomedical sciences.

Assist decision-making in public health management: LLMs have emerged as potent instruments for data-driven decision-making in public healthcare management. By analyzing and synthesizing diverse data sources, LLMs can offer comprehensive insights into disease trends, treatment efficacy and patient outcomes. These data-driven analyses aid in formulating evidence-based medical protocols and clinical guidelines, ultimately improving the overall quality of healthcare delivery (Arora & Arora, 2023). Moreover, LLMs can support public health authorities in tracking and managing infectious disease outbreaks, enabling proactive interventions to curb their spread. The integration of LLMs in public health initiatives holds the potential to revolutionize data surveillance, epidemiological modeling and health policy formulation. LLMs can process vast amounts of public health data, including health surveys, population health records and disease surveillance reports, allowing for real-time tracking of health indicators and early identification of health risks. Through predictive modeling, LLMs can assist in forecasting disease outbreaks and resource allocation, helping public health authorities prepare timely and effective responses.

While LLMs hold great promise in the fields of medical and public health domains, there remain several challenges that require thorough and thoughtful consideration. Data privacy and security concerns are paramount, as the use of LLMs entails handling sensitive patient information. Furthermore, potential biases in the generated content must be acknowledged and addressed to ensure equitable healthcare practices (Li *et al.*, 2023; Shen *et al.*, 2023). Ethical guidelines must be established to govern the use of LLMs in medical decision-making, research and public health initiatives, upholding the principles of patient autonomy, beneficence and non-maleficence. Transparent and explainable AI models are essential to foster trust among healthcare professionals, patients and the public. Additionally, the

potential displacement of certain healthcare tasks by LLMs should be carefully managed, striking a balance between human expertise and AI assistance.

5.4 Speech and Recognition Techniques

LLMs-based end-to-end speech recognition model: The emergence of LLMs like ChatGPT also benefits the speech recognition field, which is a task to map audio inputs to text outputs. Traditional speech recognition systems usually combine various components, such as a language model, an acoustic model, a pronunciation model, etc. This structure relies on large computation storage and restricts the deployment of speech recognition systems on-device implementation. Research on end-to-end speech recognition models based on the LLMs is needed to mitigate this issue. For example, end-to-end speech recognition models were reviewed and compared on recognition accuracy, latency, mode size and computational cost (Kim *et al.*, 2020).

Improve LLMs' decoding speed in speech recognition: The decoding speed of the LLMs is often slow in speech recognition, which affects the use of speech recognition in real-time applications, such as voice assistants, transcription services, communication tools, etc. Taking advantage of the Central Processing Unit (CPU) and Graphics Processing Unit (GPU) is a method to accelerate the decoding speed. By using optimized deep learning frameworks that fully leverage hardware acceleration, it is possible to speed up the computations with minimal loss in accuracy. For example, an implementation using multicore CPUs and GPUs was proposed to reduce the time of speech recognition (Kim & Lane, 2014).

Multi-modal LLMs in speech recognition: Another technique path to developing LLMs for speech recognition is considering the fusion of a text-based and speech-based LLM. The speech-based LLMs are usually good at preserving the speaker's identity information and intonation and the text-based LLMs are better than speech-based LLMs in learning linguistics knowledge. Combining both types of LLMs allows the system to leverage their respective strengths, leading to a more comprehensive understanding of the input. Recently, a combination of AudioPaLM (a speech-based language model) and PaLM-2 (a text-based language model) was proposed to improve speech tasks (Rubenstein *et al.*, 2023).

System combination for translation: Machine translation (Brants, Popat, Xu, Och, & Dean, 2007) describes the problem of translating a source-language (e.g. Chinese) sentence or audio to a target language (e.g. English) sentence or audio. The developments of LLMs are revolutionizing the field of translation as existing LLMs perform quite well on various NLP tasks, and machine translation is one of the essential natural language tasks. Typically, the decoder component of an LLM will help to generate translated outputs automatically and the LLM will be combined with other techniques to build the machine translation system. For example, a combination of the neural probabilistic language model and noise-contrastive estimation was used for a machine system (Vaswani *et al.*, 2013). The results demonstrate that combination systems can generate good-quality translations without repeating summarizations over the whole vocabulary. Moreover, the system combination method is usually used to build machine translation systems to improve the confidence of machine translation results and mitigate the ambiguity issue of natural language. It combines outputs from various translation systems. For example, if most systems contain a set of words, then the combination output would contain them in a large probability. However, doing system combinations can also be challenging as the order of the words translated by each system could differ. To improve the efficiency of the system combination in translation, a confusion network generation method (Karakos & Khudanpur, 2008) is widely used in this field.

Prompting strategies of LLMs in translation: The prompting strategies are recently explored for machine translation (Zhang, Haddow, & Birch, 2023). The selection of prompt examples could affect the LLM's performance in machine translation, like how a prompting strategy could affect the performance of an LLM. Therefore, a good prompting strategy is essential for this field. The

language of prompt examples is also important. For example, an English template or prompt could work best for machine translation than other languages. One reason for this result could be that the LLM is pretrained on English datasets. Therefore, the translation performance of LLMs could not be robust or stable when translating between German and Chinese. So, there is still a gap for LLMs doing machine translation in non-English-centric tasks.

LLMs in translation with limited storage: Although the performance of LLMs on machine translation is good, their storage requirements are also huge. Therefore, how we can efficiently train the LLMs for machine translation could be a key question for this field. Recently, there have been some works proposed to answer this question. For example, an efficient implementation of the continuous space language model (Schwenk *et al.*, 2012) was developed to accelerate the training of the language model by taking advantage of modern multicore computers. The data selection technique is also proven helpful in resampling training data in large corpora. LLMs are usually enormous because of the different combinations of source and target phrases. To keep the usage of LLMs in translation under a limited computing capacity, it is meaningful to do research on reducing the scale of model without sacrificing the translation quality. One possible method is to use suitable pruning techniques for the LLMs. For example, divergence-based fine pruning (Kim, Park, Shin, Kwon, & Kim, 2017) was used to reduce the model size of LLM for translation tasks.

6. Future research

Although LLMs such as ChatGPT show remarkable results when interacting with human prompts, they can still generate wrong responses sometimes, especially in cases they have never seen in the training datasets. Their responses' accuracy should still be improved for human usage of good confidence. Therefore, researchers need to focus on the methodology of enhancing the performance of LLMs.

On the other hand, there are potential biases in the generated results by LLMs as a vast and biased dataset is used to train them. So unbiased LLMs should be developed to create fair results for humans. There are existing unbiased methods for building machine learning models (Breeden & Leonova, 2021). Researchers need to consider how to combine or adapt these unbiased methods for LLMs to make sure they can generate accurate and unbiased results.

Moreover, deep learning models are usually considered black boxes because of their complex structures and lack of transparency and explainability. LLMs are more complex as they usually need a larger model and dataset size. Therefore, it becomes harder to understand why they are generating the responses they are. When LLMs are used for making high-stake decisions in finance, law, healthcare, etc., we need to understand better their results to increase our trust in them. Researchers in the explainable artificial intelligence (XAI) field are working very hard to solve the above issues, and there are existing XAI methods in the literature (Linardatos, Papastefanopoulos, & Kotsiantis, 2021). Therefore, combining XAI and LLMs is another significant research direction to improve the transparency of LLMs.

Finally, as LLMs gain widespread adoption across diverse domains, concerns related to ethics and privacy become critical considerations. Researchers must collaborate closely with policymakers and legal experts to develop robust frameworks that govern the ethical usage and handling of data by these models. The establishment of clear guidelines and standards ensures that LLMs are deployed responsibly, preserving users' rights and minimizing potential risks. By adhering to these principles and incorporating privacy-preserving techniques, LLMs can engender trust and confidence among users, paving the way for their responsible and beneficial integration in education, research and other domains.

To build an efficient and safe environment for LLMs, efforts should come from researchers in diverse domains. Computer scientists can help enhance LLMs' performance and implement efficient pipelines. Mathematicians and physicists can help to uncover the mechanisms

behind LLMs. Educators or healthcare professionals can provide domain knowledge when building domain-specific LLMs in their domains. Policymakers and legal experts can design the legal framework to address ethical concerns of using LLMs.

References

Angelov, D. (2020). Top2Vec: Distributed representations of topics. doi: 10.48550/arXiv.2008.09470.

Arora, A., & Arora, A. (2023). The promise of large language models in health care. *The Lancet*, *401*(10377), 641.

Atzeni, D., Bacciu, D., Mazzei, D., & Prencipe, G. (2022). A systematic review of wi-fi and machine learning integration with topic modeling techniques. *Sensors (Basel, Switzerland)*, *22*(13), 4925. doi: 10.3390/s22134925.

Bafna, P., Pramod, D., & Vaidya, A. (2016, March). Document clustering: TF-IDF approach. In *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)* (pp. 61-66). IEEE.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., … McKinnon, C. (2022). Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*.

Bianchi, F., Terragni, S., & Hovy, D. (2021). Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. *ACL-IJCNLP*, *2021*, 2021, /08//.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 10.

Bonner, E., Lege, R., & Frazier, E. (2023). LARGE LANGUAGE model-based artificial intelligence in the language classroom: Practical ideas for teaching. *Teaching English with Technology*, *23*(1), 23–41.

Brants, T., Popat, A. C., Xu, P., Och, F. J., & Dean, J. (2007). Large language models in machine translation.

Breeden, J. L., & Leonova, E. (2021). Creating unbiased machine learning models by design. *Journal of Risk and Financial Management*, *14*(11), 565. Available from: https://www.mdpi.com/1911-8074/14/11/565

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., … Askell, A. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Chang, Y., Wang, X., Wang, J., Wu, Y., Zhu, K., Chen, H., … Wang, Y. (2023). A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109*.

Chaudhry, M. A., Cukurova, M., & Luckin, R. (2022). A transparency index framework for AI in education. In *International Conference on Artificial Intelligence in Education*.

Chechitelli, A. (2023). AI writing detection update from Turnitin's chief product officer. *Turnitin Blog*.

Condor, A., Litster, M., & Pardos, Z. (2021). Automatic short answer grading with SBERT on out-of-sample questions. In *Proceedings of the 14th International Conference on Educational Data Mining (EDM2021)* (pp. 345-352).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-Training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT 2019 Jun 2* (Vol. 1, p. 2).

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd* (Vol. 96, No. 34, pp. 226-231).

Fan, L., Li, L., Ma, Z., Lee, S., Yu, H., & Hemphill, L. (2023). A bibliometric review of large language models research from 2017 to 2023. *arXiv preprint arXiv:2304.02020*.

Févotte, C., & Idier, J. (2011). Algorithms for nonnegative Matrix factorization with the β-divergence. *Neural Computation*, *23*(9), 2421–2456. doi: 10.1162/NECO_a_00168.

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. doi: 10.48550/arXiv.2203.05794.

Hsu, Y.-C., & Ching, Y.-H. (2023). Generative artificial intelligence in education, Part One: The dynamic frontier. *TechTrends*, *67*, 603–607.

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2065), 20150202.

Karakos, D., & Khudanpur, S. (2008). Sequential system combination for machine translation of speech. In *2008 IEEE Spoken Language Technology Workshop*.

Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., … Hüllermeier, E. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, *103*, 102274.

Kim, J., & Lane, I. (2014). *Accelerating large vocabulary continuous speech recognition on heterogeneous cpu-gpu platforms*. In 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 3291-3295). IEEE.

Kim, K., Park, E.-J., Shin, J.-H., Kwon, O.-W., & Kim, Y.-K. (2017). Divergence-based fine pruning of phrase-based statistical translation model. *Computer Speech and Language*, *41*, 146–160.

Kim, C., Gowda, D., Lee, D., Kim, J., Kumar, A., Kim, S., … Han, C. (2020). A review of on-device fully neural end-to-end automatic speech recognition algorithms. In *2020 54th Asilomar Conference on Signals, Systems, and Computers*.

Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., … Maningo, J. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLoS Digital Health*, *2*(2), e0000198.

Lee, H. (2023). *The rise of ChatGPT: Exploring its potential in medical education*. Anatomical Sciences Education.

Leinonen, J., Denny, P., MacNeil, S., Sarsa, S., Bernstein, S., Kim, J., … Hellas, A. (2023). Comparing code explanations created by students and large language models. *arXiv preprint arXiv:2304.03938*.

Li, H., Moon, J. T., Purkayastha, S., Celi, L. A., Trivedi, H., & Gichoya, J. W. (2023). Ethics of large language models in medicine and medical research. *The Lancet Digital Health*, *5*(6), e333–e335.

Liévin, V., Hother, C. E., & Winther, O. (2022). Can large language models reason about medical questions?. *arXiv preprint arXiv:2207.08143*.

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2021). Explainable AI: A review of machine learning interpretability methods. *Entropy*, *23*(1), 18. Available from: https://www.mdpi.com/1099-4300/23/1/18

Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., … Liu, Z. (2023). Summary of chatgpt/gpt-4 research and perspective towards the future of large language models. *arXiv preprint arXiv:2304.01852*.

Maaten, L. V. D., & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, *9*(86), 2579–2605. Available from: http://jmlr.org/papers/v9/vandermaaten08a.html

MacNeil, S., Tran, A., Leinonen, J., Denny, P., Kim, J., Hellas, A., … Sarsa, S. (2022). Automatically generating CS learning materials with Large Language Models. *arXiv preprint arXiv:2212.05113*.

Mayer, C. W., Ludwig, S., & Brandt, S. (2023). Prompt text classifications with transformer models! An exemplary introduction to prompt-based learning with large language models. *Journal of Research on Technology in Education*, *55*(1), 125–141.

McInnes, L., Healy, J., & Astels, S. (2017). hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, *2*(11), 205.

McInnes, L., Healy, J., & Melville, J. (2020). Umap: Uniform manifold approximation and projection for dimension reduction. doi: 10.48550/arXiv.1802.03426.

OpenAI (2023). GPT-4 technical report. *ArXiv./abs/2303.08774*.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., … Ray, A. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, 27730–27744.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, *1*(8), 9.

Raju Sangaraju, V., Bolla, B. K., Nayak, D. K., & Kh, J. (2022). Topic modelling on consumer financial protection bureau data: An approach using BERT based embeddings. *arXiv e-prints, arXiv-2205*.

Rubenstein, P. K., Asawaroengchai, C., Nguyen, D. D., Bapna, A., Borsos, Z., Quitry, F. D. C., … Muckenhirn, H. (2023). AudioPaLM: A large language model that can speak and listen. arXiv preprint arXiv:2306.12925.

Sallam, M. (2023). The utility of ChatGPT as an example of large language models in healthcare education, research and practice: Systematic review on the future perspectives and potential limitations. *medRxiv*, 2023.2002. 2019.23286155.

Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, *45*(11), 2673–2681.

Schwenk, H., Rousseau, A., & Attik, M. (2012). Large, pruned or continuous space language models on a gpu for statistical machine translation. In *Proceedings of the NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT*.

Shen, Y., Heacock, L., Elias, J., Hentel, K. D., Reig, B., Shih, G., & Moy, L. (2023). *ChatGPT and other large language models are double-edged swords*. In (Vol. 307, p. e230163): Radiological Society of North America.

Teubner, T., Flath, C. M., Weinhardt, C., van der Aalst, W., & Hinz, O. (2023). Welcome to the era of chatgpt et al. the prospects of large language models. *Business and Information Systems Engineering*, *65*(2), 95–101.

Ufuk, F. (2023). The role and limitations of large language models such as ChatGPT in clinical settings and medical journalism. *Radiology*, *307*(3), e230276.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., … Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, (Vol. 30)..

Vaswani, A., Zhao, Y., Fossum, V., & Chiang, D. (2013). Decoding with large-scale neural language models improves translation. In *Proceedings of the 2013 conference on empirical methods in natural language processing*.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., … Metzler, D. (2022a). Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., … Zhou, D. (2022b). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, *35*, 24824–24837.

Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., … Gašević, D. (2023). Practical and ethical challenges of large language models in education: A systematic scoping review. *arXiv preprint arXiv:2303.13379*.

Zhang, B., Haddow, B., & Birch, A. (2023). Prompting large language model for machine translation: A case study. *arXiv preprint arXiv:2301.07069*.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., … Dong, Z. (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Corresponding author
Chong Guan can be contacted at: guanchong@suss.edu.sg