information

MDPI

Article

Generative Pre-Trained Transformer (GPT) in Research: A Systematic Review on Data Augmentation

Fahim Sufi

School of Public Health and Preventive Medicine, Monash University, 553 St. Kilda Rd., Melbourne, VIC 3004, Australia; research@fahimsufi.com

Abstract: GPT (Generative Pre-trained Transformer) represents advanced language models that have significantly reshaped the academic writing landscape. These sophisticated language models offer invaluable support throughout all phases of research work, facilitating idea generation, enhancing drafting processes, and overcoming challenges like writer's block. Their capabilities extend beyond conventional applications, contributing to critical analysis, data augmentation, and research design, thereby elevating the efficiency and quality of scholarly endeavors. Strategically narrowing its focus, this review explores alternative dimensions of GPT and LLM applications, specifically data augmentation and the generation of synthetic data for research. Employing a meticulous examination of 412 scholarly works, it distills a selection of 77 contributions addressing three critical research questions: (1) GPT on Generating Research data, (2) GPT on Data Analysis, and (3) GPT on Research Design. The systematic literature review adeptly highlights the central focus on data augmentation, encapsulating 48 pertinent scholarly contributions, and extends to the proactive role of GPT in critical analysis of research data and shaping research design. Pioneering a comprehensive classification framework for "GPT's use on Research Data", the study classifies existing literature into six categories and 14 sub-categories, providing profound insights into the multifaceted applications of GPT in research data. This study meticulously compares 54 pieces of literature, evaluating research domains, methodologies, and advantages and disadvantages, providing scholars with profound insights crucial for the seamless integration of GPT across diverse phases of their scholarly pursuits.

Keywords: LLM; GPT; systematic literature review; GPT in research; data augmentation; feature extraction; synthetic data generation

Citation: Sufi, F. Generative Pre-Trained Transformer (GPT) in Research: A Systematic Review on Data Augmentation. *Information* **2024**, *15*, 99. https://doi.org/10.3390/info15020099

Academic Editors: Emilio Matricciani, Heming Jia and Zhigang Chu

Received: 22 January 2024
Revised: 5 February 2024
Accepted: 6 February 2024
Published: 8 February 2024

1. Introduction

The advent of advanced language models, particularly those exemplified by GPT (Generative Pre-trained Transformer) and LLMs (Large Language Models), such as GPT-3, has profoundly influenced the landscape of academic writing. These technologies have demonstrated considerable utility in the realm of scholarly endeavors, providing valuable support in idea generation, drafting processes, and surmounting challenges associated with writer's block [1–3]. However, a comprehensive understanding of their implications in the academic context necessitates an acknowledgment of the nuanced interplay between their benefits and limitations, as evidenced by scholarly investigations [4,5]. The scholarly discourse on GPT and LLMs reveals a dichotomy wherein their application in academic writing is accompanied by notable advantages and inherent challenges [1–3]. Noteworthy studies delve into the intricate dynamics of human–machine interaction, emphasizing the imperative of judiciously integrating AI tools into the fabric of writing practices [4]. Furthermore, recent contributions extend the conversation to encompass copywriting, elucidating the multifaceted impact of AI on diverse professional roles and creative processes [5]. Thus, while these technologies offer promising prospects for enhancing research writing, their conscientious and responsible utilization becomes paramount.

The primary challenges identified in recent scholarship pertaining to the utilization of GPT and LLMs in research writing converge on concerns related to accuracy, potential biases, and ethical considerations [6]. Addressing these challenges requires a concerted effort to establish ethical guidelines and norms, ensuring the judicious use of LLMs in research endeavors [6]. The academic discourse underscores the significance of upholding scientific rigor and transparency, particularly in light of the potential biases embedded in LLM outputs [3–5]. Papers in [7–9] collectively suggest that while LLMs offer innovative tools for research writing, their use must be accompanied by careful consideration of ethical standards, methodological rigor, and the mitigation of biases. As highlighted in [7–9], one of the darning issues of using GPT or LLM-based technology in authoring academic publications involves the use of AI-based paraphrasing to hide potential plagiarism in scientific publications.

Notwithstanding the concerns associated with the authoring aspect of research, the review at hand strategically narrows its focus to explore alternative dimensions of GPT and LLM applications in scholarly pursuits. Specifically, the examination focuses on data augmentation, where GPT and LLMs play a pivotal role in enhancing research data, generating features, and synthesizing data [10–12]. As shown in Figure 1, with GPT's advanced language understanding capabilities, features can be extracted from plain text information. It should be noted that previously, feature extraction from plain texts involved various natural language processing techniques like entity recognition, sentiment analysis, classification, and others, as shown in [13–18]. With the introduction of GPT and associated technologies, a simple prompt can extract various features from plain text (Figure 1). Moreover, as depicted in Figure 1. Semantically similar content could be added by GPT being part of the data augmentation process, improving the diversity and robustness of the data. Furthermore, rows of data could be synthetically generated by GPT, facilitating the training of the machine learning process during times of data scarcity or confidentiality (shown in Figure 1).

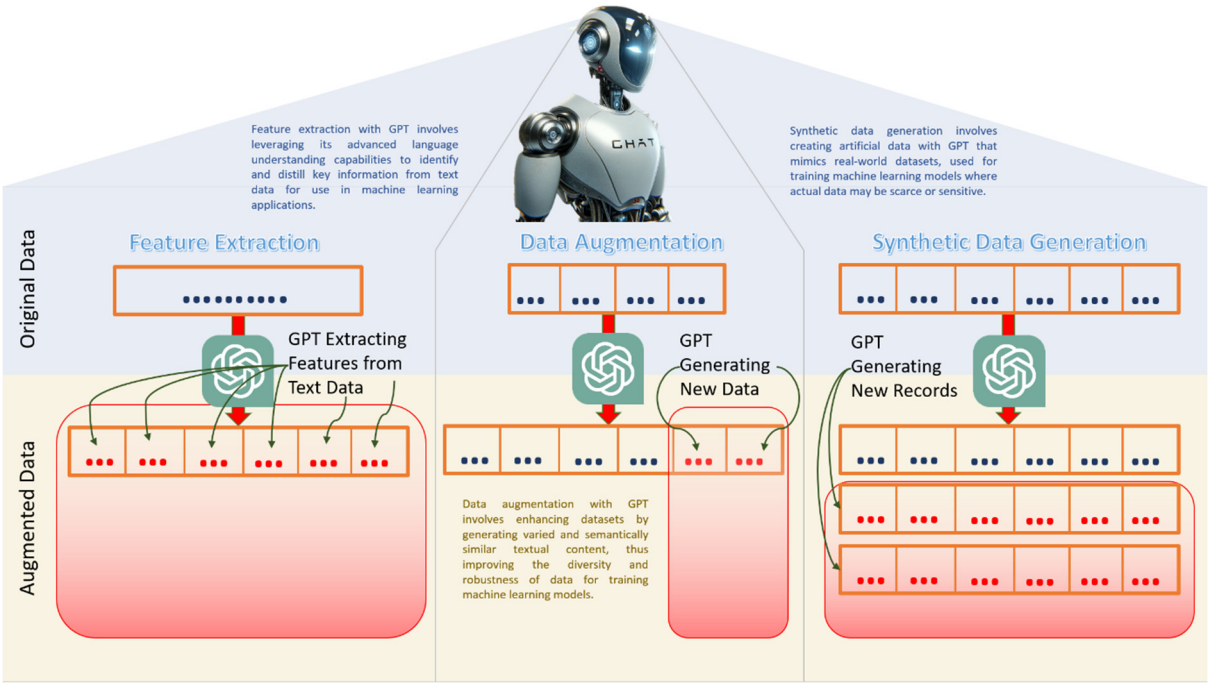

Figure 1. Conceptual diagram of how GPT performs feature extraction, data augmentation, and synthetic data generation.

In the recent literature landscape, various reviews have emerged concerning the adoption of GPT in research [19–21]. However, a critical research gap persists, as none of these reviews comprehensively explores GPT's substantial capacity and capability in the realms

of generating, processing, and analyzing research data. To address this significant void, this literature review meticulously scrutinizes 412 scholarly works, employing rigorous exclusion criteria to distill a curated selection of 77 research contributions. These selected studies specifically address three pivotal research questions, delineated as follows:

- RQ1: How can GPT and associated technology assist in generating and processing Research Data
- RQ2: How can GPT and associated technology assist in analyzing Research Data
- RQ3: How can GPT and associated technology assist in Research Design and problem solving

In Figure 2, the systematic literature review adeptly highlights the central focus on data augmentation with GPT, encapsulating 45 highly pertinent scholarly contributions. Beyond this primary focus, the study delves into the proactive role of GPT in facilitating critical analysis of research data [22–25] and shaping research design [26–28]. Significantly advancing the scholarly discourse, this study pioneers the development of a comprehensive classification framework for "GPT's use of research data", marking a seminal contribution. By critically scrutinizing existing research works on data augmentation, the review uniquely establishes a classification system encompassing six overarching categories and 14 sub-categories, providing a systematic and insightful perspective on the multifaceted applications of GPT in research data. Furthermore, the judicious placement of all 45 seminal works within these meticulously defined sub-categories serves as a sagacious validation of the intellectual rigor and innovation inherent in this classification framework, thereby substantiating its rationale and scholarly significance. Ultimately, the meticulous comparative analysis of 54 extant literary works, rigorously evaluating research domains, methodological approaches, and attendant advantages and disadvantages, provides profound insights to scientists and researchers contemplating the seamless integration of GPT across diverse phases of their scholarly endeavors.

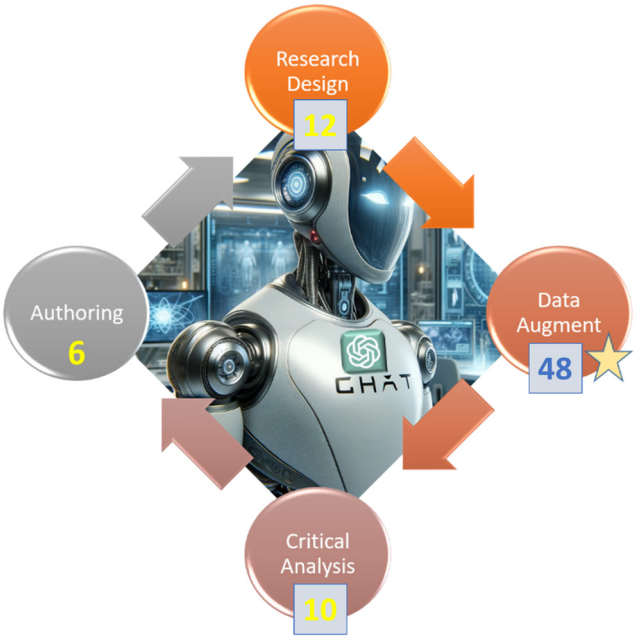

Figure 2. Use GPT and associated LLM in all phases of research. 48 scholarly works on data augmentation (Starred denoting the main focus of this review), 12 existing publications on critical analysis (i.e., research data analysis), and 10 papers on research design.

2. Research Methods

This systematic literature review rigorously investigates the application of GPT, LLM, and related technologies across various research phases following the PRISMA method, an acronym for Preferred Reporting Items for Systematic Reviews and Meta-Analyses.

PRISMA represents a meticulously structured approach for conducting and reporting systematic literature reviews and meta-analyses within the realm of academic research. The PRISMA method encompasses key components such as the identification, screening, eligibility assessment, and inclusion of pertinent studies, as well as the extraction and synthesis of data from the selected literature. This systematic approach is underpinned by a commitment to reducing bias and enhancing the reproducibility of the review process. Employing a meticulously devised strategy, as depicted in Figure 3, we formulated a comprehensive set of search keywords. These keywords were then utilized across multiple databases, including IEEE Xplore, Scopus, ACM Library, Web of Science, and PubMed, leading to the initial identification using their supported advanced queries. As seen from Figures A1–A5 in Appendix A, the actual implementation of advanced queries varies from platform to platform. However, all these advanced queries implemented the conceptual query design represented in Figure 3. Additionally, we explored other sources like Litmaps [29], yielding nine supplementary resources. Innovative visualization tools like Litmaps showed related research in our domain of interest, thereby highlighting papers that might have been missed from mainstream databases. Figure A6 of Appendix A shows how Litmaps identified 20 possible citations that might be within the domain of interest. After identification of possible records from IEEE Xplore, Scopus, ACM Library, Web of Science, and PubMed, duplicated records were identified and removed. From the records without duplicates, the screening process was performed in two stages. In the first stage, screening was performed by careful inspection of the titles and abstracts. If the title and abstract of a record contained the keywords but focused on a completely different area of research, then that record was excluded. For example, the records shown in Figure A1 (i.e., GPT as virtual assistant in medical surgery), Figure A2 (i.e., comparison of ChatGPT, GPT, and DALL-E2), and Figure A4 (i.e., research in nanomaterials and nanotechnology) of Appendix A were excluded because these studies did not focus on "GPT in Research". In stage 2 of screening, full-text articles were downloaded, inspected, and critically reviewed for eligibility. From reading the full text, if the article had an insignificant focus on the research questions set out in RQ1, RQ2, and RQ2, the article was excluded. Finally, 77 articles relevant to the research questions were included in this study.

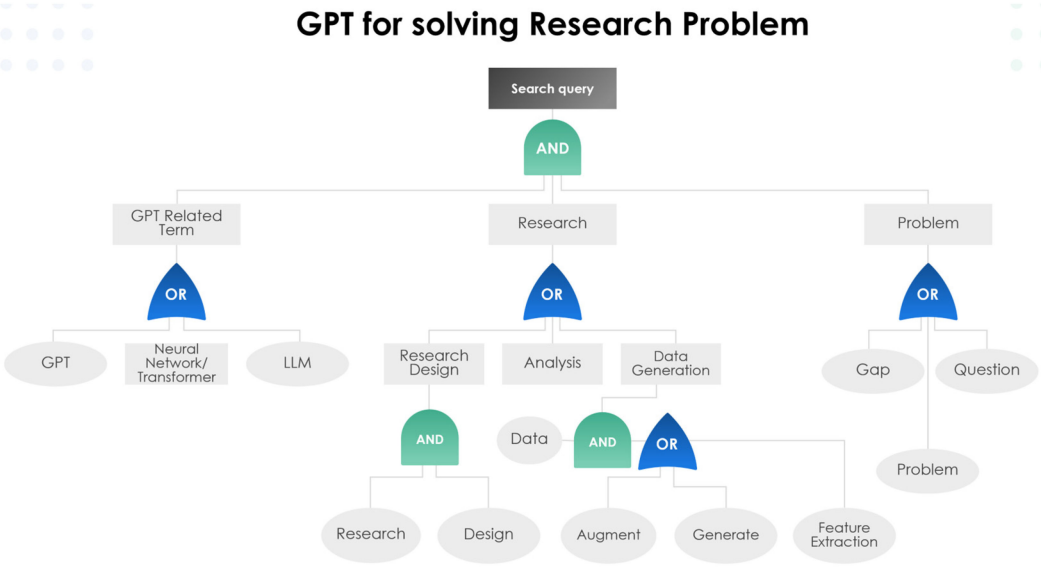

Figure 3. Search keyword used for obtaining relevant existing academic works on "GPT, LLM, and associated technologies in different phases of research".

It should be noted that the extensive set of criteria represented in Table 1 is a modified version of an earlier work on systematic literature review [30]. Figure 4 presents a schematic diagram elucidating this systematic review process. Our analysis reveals that 48 of these

papers directly address GPT/LLM applications in research data phases, while another 22 contribute insights on their utilization in research analysis and design.

Table 1. Inclusion and Exclusion criteria for both peer-reviewed and grey literature.

	Category	Criteria
Inclusion	Peer Reviewed Literature	• (GPT OR LLM) AND (Research Design OR (DATA AND (Augment OR Generate OR Feature Extraction)) OR Data Generate) AND (GAP OR Question OR Problem) • Review studies, Survey/Questionnaire based qualitative or quantitative studies, Original Research articles • Paper indexed in popular peer-reviewed sources (i.e., IEEE Explore, ACM Library, PubMed, Scopus, and Web of Science) • Papers focusing into research questions RQ 1–RQ 3 • Studies available in English language • Studies available in full text
	Grey Literature	• Websites focused into low code development platforms and their features • Indexed in popular search engines (i.e., Google and Microsoft Bing) • Articles authored by either by the GPT/LLM vendor or third-party benchmarking company • Article available in English language
Exclusion	Peer Reviewed Literature	• Tutorial Papers • Short papers less than four pages • Poster Papers, Editorials, Abstract (i.e., lacking detailed Information) • Papers prior to 2019 (Since GPT was introduced in 2019)
	Grey Literature	• Websites referring to peer reviewed literature • GPT Platforms promoted by bloggers, consultants or third-party companies • Tutorial Videos and discussions on GPT

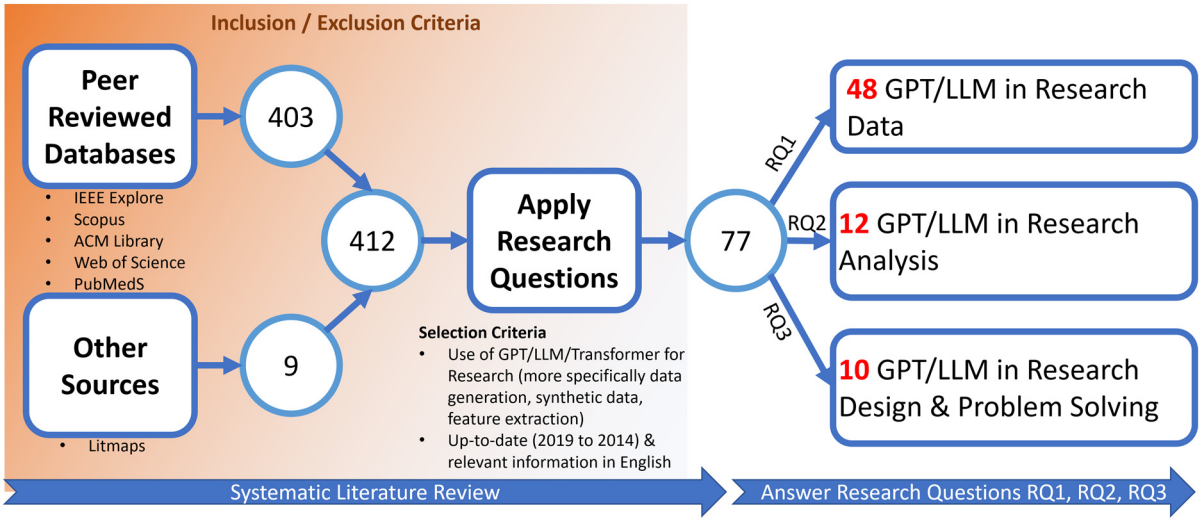

Figure 4. Schematic Diagram of the systematic literature review (i.e., use of GPT, LLM, and associated technologies on different phases of research).

Significantly, a subset of 6+ papers [3–9] underscore major concerns associated with GPT-based technologies in research authorship, citing issues such as hallucinations, biases, ethical dilemmas, and the potential for AI-assisted plagiarism concealment. Consequently, this review conscientiously excludes GPT-based research authoring from its primary focus areas.

3. Existing Research on GPT's Use in Research Data

As shown in Figures 2 and 4, the major focus of this review study is on the use of GPT in research data. As a result of the systematic literature review process, 48 existing pieces of literature were found to be highly relevant to "GPT on Research Data". This section would summarize these papers.

GPT and other large language models (LLMs) provide versatile tools for various data-related tasks. They excel at generating coherent, contextually relevant textual data, making them ideal for content creation across diverse fields. LLMs can synthesize realistic synthetic data, which is especially valuable in domains with privacy concerns or data scarcity. In data augmentation, these models enhance existing datasets by adding new, synthesized samples, thereby improving the robustness of machine learning models. Furthermore, LLMs are capable of extracting and generating features from complex datasets, aiding in more efficient and insightful data analysis. The adaptability of these models to different data types and their ability to tailor their output make them powerful tools in data science and AI development.

The study in [31] presents the GReaT (Generation of Realistic Tabular data) approach, which uses transformer-based large language models (LLMs) for the generation of synthetic tabular data. This method addresses challenges in tabular data generation by leveraging the generative capabilities of LLMs. [32] discusses a method that combines GPT technology with blockchain technology for secure and decentralized data generation. This approach can be beneficial in scenarios where data privacy and security are paramount. [26] explores the use of GPT for augmenting existing datasets. It demonstrates how GPT can be used to expand datasets by generating new, realistic samples, which can be particularly useful in fields where data is scarce or expensive to obtain. A study in [33] presents a novel application of GPT for feature extraction from unstructured data. It showcases how GPT models can be fine-tuned to identify and extract relevant features from complex datasets, enhancing data analysis and machine learning tasks. Ref. [34] details an early version of a GPT-based system for data anonymization, highlighting its potential to protect sensitive information in datasets while retaining its utility for analysis. Research in [35] focuses on using GPT for generating synthetic datasets for training machine learning models. This is especially useful in domains where real-world data is limited or sensitive. The work in [36] investigates the potential of Foundation Models (FMs), like GPT-3, in handling classical data tasks such as data cleaning and integration. It explores the applicability of these models to various data tasks, demonstrating that large FMs can adapt to these tasks with minimal or no task-specific fine-tuning, achieving state-of-the-art performance in certain cases. The study in [37] discusses advanced techniques for text data augmentation, focusing on improving the diversity and quality of the generated text data. It emphasizes methods that can generate nuanced and contextually appropriate data, enhancing the performance of machine learning models, particularly in natural language processing tasks. Ref. [38] explores the use of GPT-based models for feature extraction and generation, showcasing their effectiveness in identifying and creating relevant features from complex datasets, which is critical for enhancing data analysis and predictive modeling.

The work in [39] introduces LAMBADA, a novel method for text data augmentation. It leverages language models for synthesizing labeled data to improve text classification tasks, particularly in scenarios with limited labeled data. Ref. [40] focuses on the augmentation of medical datasets using transformer-based text generation models. It particularly addresses the challenge of data scarcity in the medical domain by generating synthetic clinical notes, which are then evaluated for their utility in downstream NLP tasks like unplanned readmission prediction and phenotype classification. A study in [41] presents a method called PREDATOR for text data augmentation, which improves the quality of textual datasets through the synthesis of new, high-quality text samples. This method is particularly useful for text classification tasks and demonstrates a significant improvement in model performance. Work in [42] presents a novel approach for enhancing hate speech detection on social networks. It combines DeBERTa models with back-translation and

GPT-3 augmentation techniques during both training and testing. This method significantly improves hate speech detection across various datasets and metrics, demonstrating robust and accurate results. The work in [43] presents DHQDA, a novel method for data augmentation in Named Entity Recognition (NER). It uses GPT and a small-scale neural network for prompt-based data generation, producing diverse and high-quality augmented data. This approach enhances NER performance across different languages and datasets, particularly in low-resource scenarios. The study in [44] introduces the I-WAS method, a data-augmentation approach using GPT-2 for simile detection. It focuses on generating diverse simile sentences through iterative word replacement and sentence completion, significantly enhancing simile detection capabilities in natural language processing. Ref. [45] explores generation-based data augmentation for offensive language detection, focusing on its effectiveness and potential bias introduction. It critically analyzes the feasibility and impact of generative data augmentation in various setups, particularly addressing the balance between model performance improvement and the risk of bias amplification. Ref. [46] explores the use of GPT-2 in generating synthetic biological signals, specifically EEG and EMG, to improve classification in biomedical applications. It demonstrates that models trained on synthetic data generated by GPT-2 can achieve high accuracy in classifying real biological signals, thus addressing data scarcity issues in biomedical research. The work in the [47] document describes a study on fine-grained claim detection in financial documents. The research team from Chaoyang University of Technology uses MacBERT and RoBERTa with BiLSTM and AWD-LSTM classifiers, coupled with data resampling and GPT-2 augmentation, to address data imbalance. They demonstrate that data augmentation significantly improves prediction accuracy in financial text analysis, particularly in the Chinese Analyst's Report section. Ref. [48] discusses the role of ChatGPT in data science, emphasizing its potential in automating workflows, data cleaning, preprocessing, model training, and result interpretation. It highlights the advantages of ChatGPT's architecture, its ability to generate synthetic data, and addresses limitations and concerns such as bias and plagiarism.

Research work in [49] details an approach for automating the extraction and classification of technical requirements from complex systems' specifications. It utilizes data augmentation methods, particularly GPT-J, to generate a diverse dataset, thereby enhancing the training of AI models for better classification accuracy and efficiency in requirements engineering. Ref. [50] investigates data augmentation for hate speech classification using a single class-conditioned GPT-2 language model. It focuses on the multi-class classification of hate, abuse, and normal speech and examines how the quality and quantity of generated data impact classifier performance. The study demonstrates significant improvements in macro-averaged F1 scores on hate speech corpora using the augmented data. The study in [51] explores the use of GPT models for generating synthetic data to enhance machine learning applications. It emphasizes the creation of diverse and representative synthetic data to improve machine learning model robustness. The paper in [52] focuses on augmenting existing datasets using GPT models. The method involves generating additional data that complements the original dataset, thereby enhancing the richness and diversity of data available for machine learning training.

The research in [12] examines the generation of synthetic educational data using GPT models, specifically for physics education. It involves creating responses to physics concept tests, aiming to produce diverse and realistic student-like responses for educational research and assessment design. The reference in [53] discusses the revolutionary role of artificial intelligence (AI), particularly large language models (LLMs) like GPT-4, in generating original scientific research, including hypothesis formulation, experimental design, data generation, and manuscript preparation. The study showcases GPT-4's ability to create a novel pharmaceutics manuscript on 3D printed tablets using pharmaceutical 3D printing and selective laser sintering technologies. GPT-4 managed to generate a research hypothesis, experimental protocol, photo-realistic images of 3D printed tablets, believable analytical data, and a publication-ready manuscript in less than an hour. Ref. [54]

focuses on the application of GPT models to augment existing datasets in the field of healthcare. It presents a method for generating synthetic clinical notes in Electronic Health Records (EHRs) to predict patient outcomes, addressing challenges in healthcare data such as privacy concerns and data scarcity. The paper in [11] deals with textual data augmentation in the context of patient outcome prediction. The study introduces a novel method for generating artificial clinical notes in EHRs using GPT-2, aimed at improving patient outcome prediction, especially the 30-day readmission rate. It employs a teacher–student framework for noise control in the generated data. The study in [55] focuses on leveraging large language models (LLMs) like GPT-3 for generating synthetic data to address data scarcity in biomedical research. It discusses various strategies and applications of LLMs in synthesizing realistic and diverse datasets, highlighting their potential for enhancing research and decision-making in the biomedical field. Ref. [56] discusses the use of GPT models for enhancing data quality in the context of social science research. It focuses on generating synthetic responses to survey questionnaires, aiming to address issues of data scarcity and respondent bias.

The seminal work in [57] explores the use of GPT-3 for creating synthetic data for conversational AI applications. It evaluates the effectiveness of synthetic data in training classifiers by comparing it with real user data and analyzing semantic similarities and differences. The paper in [58] presents a system developed for SemEval-2023 Task 3, focusing on detecting genres and persuasion techniques in multilingual texts. The system combines machine translation and text generation for data augmentation. Specifically, genre detection is enhanced using synthetic texts created with GPT-3, while persuasion technique detection relies on text translation augmentation using DeepL. The approach demonstrates effectiveness by achieving top-ten rankings across all languages in genre detection, and notable ranks in persuasion technique detection. The paper outlines the system architecture utilizing DeepL and GPT-3 for data augmentation, experimental setup, and results, highlighting the strengths and limitations of the methods used. Ref. [59] explores using GPT-3 for generating synthetic responses in psychological surveys. It focuses on enhancing the diversity of responses to understand a broader range of human behaviors and emotions. The study in [60] investigates the use of GPT-3 in augmenting data for climate change research. It generates synthetic data, representing various climate scenarios, to aid in predictive modeling and analysis.

The research in [61] leverages GPT-3.5 for augmenting Dutch multi-label datasets in vaccine hesitancy monitoring. The paper discusses how synthetic tweets are generated and used to improve classification performance, especially for underrepresented classes. Romero-Sandoval et al. [62] investigate using GPT-3 for text simplification in Spanish financial texts, demonstrating effective data augmentation to improve classifier performance. Rebboud et al. [63] explore GPT-3's ability to generate synthetic data for event relation classification, enhancing system accuracy with prompt-based, manually validated synthetic sentences. Quteineh et al. [64] present a method combining GPT-2 with Monte Carlo Tree Search for textual data augmentation, significantly boosting classifier performance in active learning with small datasets. Suhaeni et al. [65] explore using GPT-3 for generating synthetic reviews to address class imbalances in sentiment analysis, specifically for Coursera course reviews. It shows how synthetic data can enhance the balance and quality of training datasets, leading to improved sentiment classification model performance. Singh et al. [66] introduce a method to augment interpretable models using large language models (LLMs). The approach, focusing on transparency and efficiency, shows that LLMs can significantly enhance the performance of linear models and decision trees in text classification tasks. Sawai et al. discuss using GPT-2 for sentence augmentation in neural machine translation. The approach aims to improve translation accuracy and robustness, especially for languages with different linguistic structures. The method demonstrated significant improvements in translation quality across various language pairs.

The paper in [10] focuses on using GPT-2 to generate complement sentences for aspect term extraction (ATE) in sentiment analysis. The study introduces a multi-step training

procedure that optimizes complement sentences to augment ATE datasets, addressing the challenge of data scarcity. This method significantly improves the performance of ATE models. A study in [67] explores data augmentation for text classification using transformer models like BERT and GPT-2. It presents four variants of augmentation, including masking words with BERT and sentence expansion with GPT-2, demonstrating their effectiveness in improving the performance of text classification models. Another recent work by Veyseh et al. focuses on enhancing open-domain event detection using GPT-2-generated synthetic data [68]. The study introduces a novel teacher–student architecture to address the noise in synthetic data and improve model performance for event detection. The experiments demonstrate significant improvements in accuracy, showcasing the effectiveness of this approach. Waisberg et al. discuss the potential of ChatGPT (GPT-3) in medicine [69]. The study highlights ChatGPT's ability to perform medical tasks like writing discharge summaries, generating images from descriptions, and triaging conditions [69]. It emphasizes the model's capacity for democratizing AI in medicine, allowing clinicians to develop AI techniques. The paper also addresses ethical concerns and the need for compliance with healthcare regulations [69].

The paper "Investigating Paraphrasing-Based Data Augmentation for Task-Oriented Dialogue Systems" by Liane Vogel and Lucie Flek explores data augmentation in task-oriented dialogue systems using paraphrasing techniques with GPT-2 and Conditional Variational Autoencoder (CVAE) models [70]. The study demonstrates how these models can effectively generate paraphrased template phrases, significantly reducing the need for manually annotated training data while maintaining or even improving the performance of a natural language understanding (NLU) system [70]. Shuohua Zhou and Yanping Zhang focus on improving medical question-answering systems [71]. It employs a combination of BERT, GPT-2, and T5-Small models, leveraging GPT-2 for question augmentation and T5-Small for topic extraction [71]. The approach demonstrates enhanced prediction accuracy, showcasing the model's potential in medical question-answering and generation tasks.

4. A New Classification Scheme: GPT on Research Data

The previous section summarized 48 existing research works on the use of GPT in generating research data. This section demonstrates the extensive classification framework developed to group these 48 existing literary works on the use of GPT for data augmentation, natural language processing-based feature extraction, machine learning model-based data cleaning, transformation, performance improvement, etc. As seen in Figure 5, research data enhancement using GPT could be grouped into six categories. Each of these six categories hosts two or more sub-categories. In this section, all these categories and sub-categories will be described. Most importantly, all 48 existing works of literature would be classified into one or more of these categories (as shown in Tables 2–7).

4.1. Data Generation and Augmentation

- Synthetic Data Creation: Focuses on using GPT models to generate artificial data that mimics real-world data, useful in scenarios where data privacy is crucial or actual data is limited. Literature in [31,35,51,52,56,57] could be attributed to this sub-category.
- Text Data Expansion and Enhancement: This involves leveraging GPT to create new textual content and enhance existing datasets, thereby improving machine learning models' performance and addressing data scarcity [12,26,37,39,41,48,53,54,65].

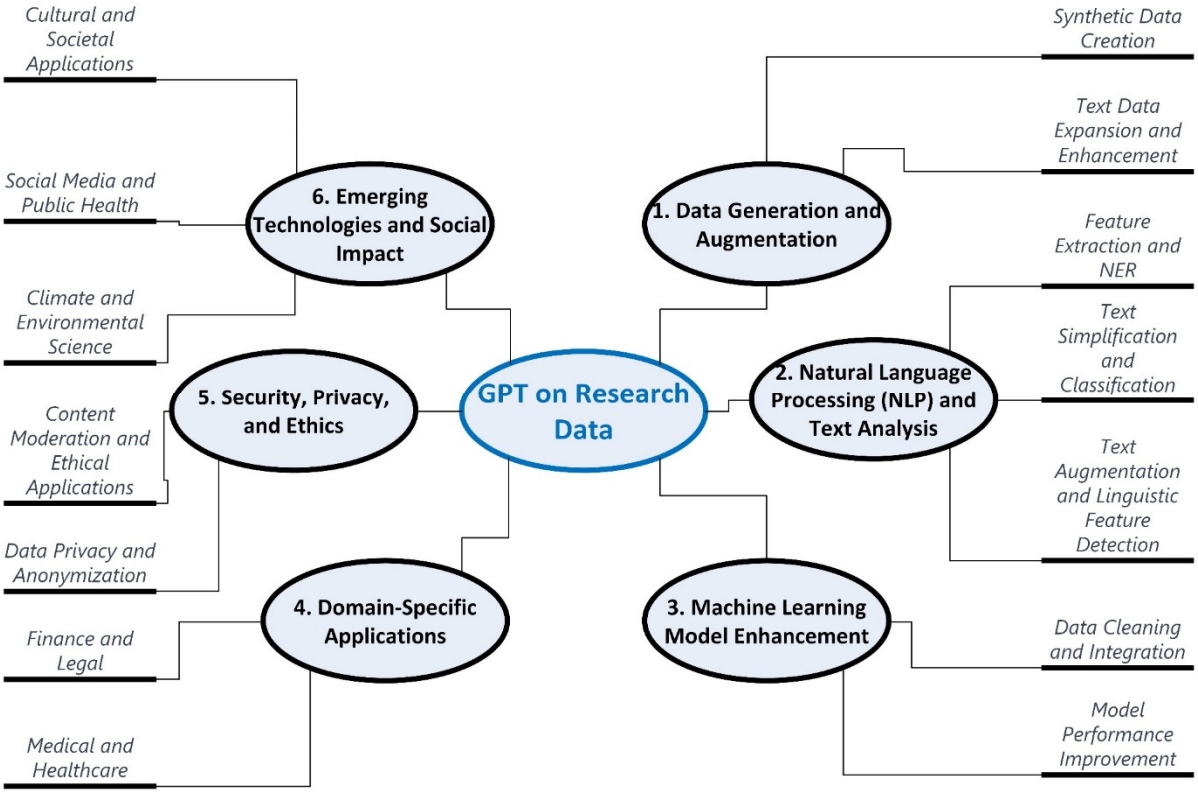

Figure 5. A comprehensive classification framework for "GPT's use of research data".

4.2. Natural Language Processing (NLP) and Text Analysis

- Feature Extraction and NER: This encompasses the use of GPT for identifying salient features within unstructured data and enhancing named entity recognition (NER) capabilities [33,38,43]. As shown in Figure 6, prior to the era of GPT, various NLP techniques like entity recognition, sentiment analysis, and category classifications were used for extracting features from textual data like social media posts, as shown in [13–18]. Then these extracted features were used by machine learning techniques for obtaining insights and analytics. However, with the advent of GPT, features could be extracted with a simple GPT prompt like "Categorize these data into the following categories (1) animals, (2) plants, and (3) equipment," as shown in Figure 5.
- Text Simplification and Classification: Covers the application of GPT in simplifying complex texts for better understanding and classification, particularly in specialized fields like finance [10,62,64,67].
- Text Augmentation and Linguistic Feature Detection: This refers to the use of GPT for generating text that aids in the detection and analysis of specific linguistic features, such as similes or event relations. Works in [44,63] (Event Relation Classification) [53] are categorized within this bucket.

Table 2. Categorizing existing literature into the "Data Generation and Augmentation" category.

Reference	Area of Research	Method Used	Advantages	Disadvantages
[31]	Realistic Tabular Data Generation	Textual Encoding and LLMs for Tabular Data	Versatile, less preprocessing, authentic synthesis	May require large models for complexity
[26]	Data Augmentation	GPT for expanding datasets	Generates new, realistic samples	May not capture specific domain nuances
[35]	Synthetic Data Generation for ML	GPT for synthetic dataset creation	Useful in limited/sensitive data scenarios	Potential biases in generated data
[37]	Text Data Augmentation	Advanced text augmentation techniques	Improves data diversity and quality	May require significant computational resources
[39]	Text Data Augmentation	LAMBADA for text data synthesis	Effective for improving text classification with limited data	May not generalize across all data types
[41]	Text Data Augmentation	PREDATOR for text synthesis	Enhances text dataset quality, useful for classification tasks	Specific focus on text data, may not generalize to other data types
[48]	Data Science Automation	ChatGPT for automating data science workflows	Streamlines data science tasks, generates synthetic data	Bias, plagiarism concerns
[12]	Educational Research, Physics Education	Generating synthetic responses for physics	Produces realistic student-like responses	Limited to specific educational contexts
[52]	Data Augmentation, Machine Learning	Augmenting datasets with GPT-generated data	Increases dataset richness; aids model training	Risk of introducing biases in synthetic data
[51]	Machine Learning, Synthetic Data Generation	Generating diverse synthetic data using GPT models	Enhances model robustness; improves data diversity	May not capture nuanced real-world scenarios
[54]	Healthcare, EHR Data Augmentation	Generating synthetic clinical notes using GPT models	Addresses healthcare data scarcity and privacy	Specific to healthcare data, may lack versatility
[56]	Social Science Research	Generating synthetic survey responses with GPT models	Addresses data scarcity and respondent bias	May not fully capture human variability
[58]	Multilingual Text Analysis	Genre and persuasion techniques detection using GPT-3 and DeepL	Effective in multilingual context; top rankings achieved	Limited by translation accuracy and text generation capabilities
[57]	Conversational AI	Synthetic data creation for AI using GPT-3	Useful for AI training; analyzes data similarities	Less effective than real user data
[65]	Sentiment Analysis with Class Imbalance	GPT-3 for generating synthetic reviews	Balances training data, improves model accuracy	
[53]	Text Augmentation and Linguistic Feature Detection generating comprehensive research documents	GPT-4 for scientific research generation	Accelerates medical research, demonstrates multidisciplinary AI capabilities	Inaccuracies in literature, experimental data challenges

Table 3. Categorizing existing literature into the "NLP and Text Analysis" category.

Reference	Area of Research	Method Used	Advantages	Disadvantages
[33]	Feature Extraction from Unstructured Data	GPT for feature identification	Efficient in complex datasets	Requires fine-tuning for specific tasks
[38]	Feature Generation and Extraction	GPT-based models for feature extraction	Effective in complex datasets	Requires model tuning for specific tasks
[43]	Named Entity Recognition (NER)	DHQDA using GPT for NER data augmentation	Enhances NER performance, effective in multiple languages	May require complex model training and fine-tuning
[44]	Simile Detection in NLP	I-WAS with GPT-2 for simile sentence generation	Enhances simile detection with diverse sentences	May require iterative adjustments and fine-tuning
[62]	Text Simplification in Spanish Financial Texts	GPT-3 for data augmentation	Effective augmentation for small datasets, improved classifier performance	Not specified
[63]	Event Relation Classification	Prompt-based GPT-3 data generation, manual validation	Realistic and relevant synthetic sentence generation, improved system performance	Manual validation process required
[64]	Text Classification with Active Learning on Small Datasets	GPT-2 with MCTS and entropy optimization	Significant performance increase in classifiers, efficient for small datasets	Potential for generation of less relevant or noisy data in outputs
[10]	Aspect Term Extraction in Sentiment Analysis	GPT-2 for generating complement sentences	Improves ATE model performance, addresses data scarcity	Complexity in optimizing sentence generation
[67]	Text Classification	Data augmentation with transformers	Enhances text classification model performance	Depends on the transformer model's limitations
[53]	Text Augmentation and Linguistic Feature Detection by generating comprehensive research documents	GPT-4 for scientific research generation	Accelerates medical research, demonstrates multidisciplinary AI capabilities	Inaccuracies in literature, experimental data challenges

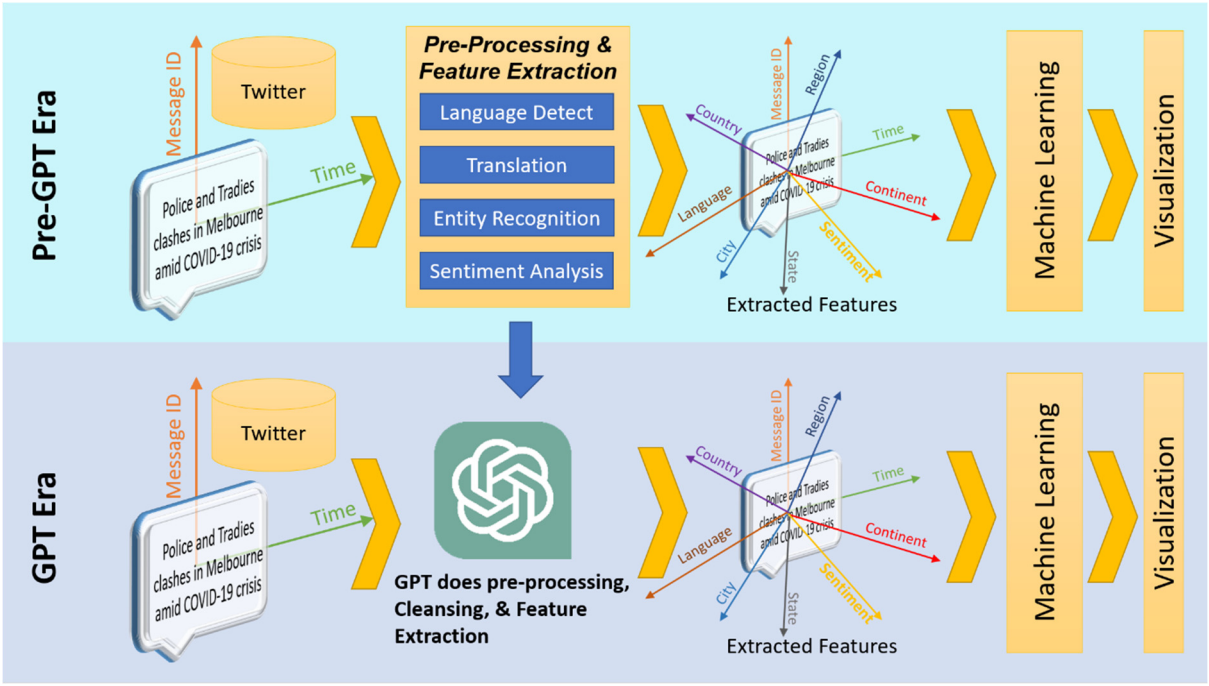

Figure 6. A comparative schematic of feature extraction process with NLP and GPT.

4.3. Machine Learning Model Enhancement

- Data Cleaning and Integration: Demonstrates how GPT could be used for data cleansing, transformation, and integration activities, as demonstrated in [36].

- Model Performance Improvement: Details how GPT and other language models are integrated during the model fitting process to enhance the performance of machine learning models, particularly those requiring interpretability [53,66,68,72].

Table 4. Categorizing existing literature into the "Machine Learning Model Enhancement" category.

Reference	Area of Research	Method Used	Advantages	Disadvantages
[36]	Data Cleaning and Integration	Foundation Models (e.g., GPT-3)	Adaptable to various data tasks with minimal fine-tuning, state-of-the-art performance in some cases	May require task-specific adaptations, complexity in handling domain-specific data
[66]	Enhancing Interpretable Models with LLMs	Aug-imodels using LLMs during model fitting	Improves interpretable models while maintaining transparency and efficiency	Requires careful integration of LLMs
[72]	Neural Machine Translation Enhancement	GPT-2 for sentence data generation	Improves translation accuracy and robustness	Time-intensive sentence generation process
[68]	Open-Domain Event Detection	GPT-2 generated synthetic data, teacher–student architecture	Significant accuracy improvements in event detection	Complex architecture and noise management
[53]	LLMs in generating coherent and contextually relevant scientific text	GPT-4 for scientific research generation	Accelerates medical research, demonstrates multidisciplinary AI capabilities	Inaccuracies in literature, experimental data challenges

4.4. Domain-Specific Applications

- Medical and Healthcare: Describes the use of GPT models for generating and augmenting data in the medical field, such as patient records, to improve healthcare outcomes and NLP tasks. Research works in [11,40,53–55,69,71,73] could be directly attributed to this sub-category.
- Finance and Legal: Discusses the application of GPT in financial and legal document analysis, aiming to improve prediction accuracy and automate document-related processes [46,47].

Table 5. Categorizing existing literature into the "Domain-Specific Applications" category.

Reference	Area of Research	Method Used	Advantages	Disadvantages
[40]	Medical Dataset Augmentation	Transformer-based text generation for clinical notes	Addresses data scarcity in medical domain, useful for NLP tasks	Quality of synthetic notes may vary
[46]	Biomedical Signal Classification	GPT-2 for generating synthetic EEG and EMG signals	Addresses data scarcity in biomedical research, high classification accuracy	Potential challenges in generating realistic biosignals
[47]	Financial Document Analysis	MacBERT, RoBERTa, BiLSTM, AWD-LSTM, GPT-2 Augmentation	Improves prediction accuracy in financial texts	Complexity in combining multiple approaches
[54]	Healthcare, EHR Data Augmentation	Generating synthetic clinical notes using GPT models	Addresses healthcare data scarcity and privacy	Specific to healthcare data, may lack versatility
[11]	Healthcare, Patient Outcome Prediction	Textual data augmentation using GPT-2 for EHRs	Improves prediction of patient outcomes	Focused on healthcare context only

Table 5. *Cont.*

Reference	Area of Research	Method Used	Advantages	Disadvantages
[55]	Biomedical Research	Synthetic data generation using large language models	Enhances biomedical research with diverse data	May face challenges in data realism and accuracy
[69]	AI in Medicine with ChatGPT	ChatGPT for medical tasks	Democratizes AI in medicine, versatile in medical tasks	Ethical concerns, regulatory compliance
[71]	Medical Question Answering	BERT, GPT-2, T5-Small for question augmentation	Enhanced prediction accuracy in medical QA	Complexity in managing multiple models
[74]	Medical and Healthcare Message Generation	AI-generated health messages using Bloom for message generation	Efficient generation of health awareness messages, potential for wide application in health communication	Dependence on the quality and biases of the underlying AI model
[53]	Medical and Healthcare, demonstrating the use of AI in accelerating research in this domain	GPT-4 for scientific research generation	Accelerates medical research, demonstrates multidisciplinary AI capabilities	Inaccuracies in literature, experimental data challenges

4.5. Security, Privacy, and Ethics

- Data Privacy and Anonymization: Addresses the use of GPT to anonymize sensitive data, striking a balance between maintaining data utility and protecting privacy [32,34].
- Content Moderation and Ethical Applications: This involves using GPT for ethical applications, such as detecting hate speech and extracting technical requirements, ensuring content moderation, and adhering to ethical standards. Research works in [42,45,49,50] (Technical Requirements Extraction) directly fall into this sub-category.

Table 6. Categorizing existing literature into the "Security, Privacy, and Ethics" category.

Reference	Area of Research	Method Used	Advantages	Disadvantages
[32]	Secure and Decentralized Data Gen.	GPT with Blockchain	Enhances data privacy and security	Complexity in implementation
[34]	Data Anonymization	GPT for anonymizing data	Protects sensitive information	Balancing data utility and anonymity is challenging
[42]	Hate Speech Detection	DeBERTa with back-translation and GPT-3	Significantly improves hate speech detection accuracy	Complexity in integrating multiple techniques
[45]	Offensive Language Detection	Generative data augmentation	Potentially improves model performance	Risk of amplifying biases
[49]	Technical Requirements Extraction	GPT-J for data augmentation in requirements engineering	Enhances classification accuracy and efficiency	Requires careful data preparation
[50]	Hate Speech Classification	Data augmentation with GPT-2	Improves classification performance in hate speech detection	Requires careful tuning to avoid bias in generated data

Table 7. Categorizing existing literature into the "Emerging Technologies and Social Impact" category.

Reference	Area of Research	Method Used	Advantages	Disadvantages
[56]	Social Science Research	Generating synthetic survey responses with GPT models	Addresses data scarcity and respondent bias	May not fully capture human variability
[60]	Climate Change Research	Generating synthetic climate reports and datasets with GPT-3	Broader understanding of climate impacts	Risk of misrepresentation and overfitting
[59]	Music Recommendation Systems	Synthetic user review and metadata generation with GPT-3	Improved recommendation accuracy and diversity	Authenticity and bias concerns in synthetic data
[61]	Vaccine Hesitancy Monitoring	Generating anti-vaccination tweets in Dutch with GPT-3.5	Improved classification of underrepresented classes	Decreased precision for common classes; generalization concerns
[70]	Task-Oriented Dialogue Systems	Paraphrasing with GPT-2 and CVAE	Reduces need for manual data, maintains/improves NLU system performance	
[58]	Multilingual Text Analysis	Genre and persuasion techniques detection using GPT-3 and DeepL	Effective in multilingual context; top rankings achieved	Limited by translation accuracy and text generation capabilities

4.6. Emerging Technologies and Social Impact

- Climate and Environmental Science: Explores the use of GPT in generating reports and datasets to broaden our understanding of climate impacts and inform environmental policies [60].
- Social Media and Public Health: Covers the generation of content like anti-vaccination tweets to monitor public health trends and sentiment, aiming for better-informed public health strategies [61].
- Cultural and Societal Applications: This includes the generation of synthetic user reviews and metadata for music recommendation systems, and the enhancement of multilingual text analysis for cultural and societal research [58,59,70].

5. Research Analysis

GPT-based technologies allow modern researchers to analyze their data with the help of prompts. A researcher can use GPT, LLM, and associated technologies for data analysis and critical research by leveraging their ability to perform complex textual analysis and pattern recognition in large datasets, aiding in tasks like detecting nuanced patterns in financial texts [23,24,75,76]. These models also excel at solving intricate problems, such as those in discrete mathematics, showing significant improvements in advanced versions like GPT-4. Furthermore, they provide innovative methodologies for analyzing statistical data, offering insights and predictions with higher efficiency compared to traditional methods. For example, the paper in [23] presents a novel system for generating data visualizations directly from natural language queries using LLMs like ChatGPT and GPT-3. The system, named Chat2VIS (as demonstrated in Figure 7), demonstrates efficient and accurate end-to-end solutions for visualizing data based on user queries. It addresses the challenge of interpreting natural language in data visualization and utilizes advanced LLMs to convert free-form natural language into appropriate visualization code. The study includes case studies and comparisons of GPT-3, Codex, and ChatGPT performances in generating visualizations from various types of queries, highlighting their potential in rendering visualizations from natural language, even when queries are ambiguous or poorly specified. The study in [24] explores the application of GPT-3 for statistical data analysis. It proposes a method for analyzing large datasets using GPT-3 to predict insights from calculated statistics. The research addresses the limitations of existing methods and

compares traditional statistical analysis with machine learning approaches using GPT-3. It includes experiments on different datasets like e-commerce sales, heart attacks, and telecom churn rates, assessing GPT-3's performance in providing insights and its accuracy compared to traditional methods. The study also discusses the pros and cons of using GPT-3 in research, focusing on performance, accuracy, and reliability.

	Title	Worldwide Gross	Production Budget	Release Year	Content Rating	Running Time	Genre	Creative Type	Rotten Tomatoes Rating	IMDB Rating
0	From Dusk Till Dawn	25728961	20000000	1996	R	107	Horror	Fantasy	63	7.1000
1	Broken Arrow	148345997	65000000	1996	R	108	Action	Contemporary Fiction	55	5.8000
2	City Hall	20278055	40000000	1996	R	111	Drama	Contemporary Fiction	55	6.1000
3	Happy Gilmore	38623460	10000000	1996	PG-13	92	Comedy	Contemporary Fiction	58	6.9000
4	Fargo	51204567	7000000	1996	R	87	Thriller	Contemporary Fiction	94	8.3000
5	The Craft	55669466	15000000	1996	R	100	Thriller	Fantasy	45	5.9000
6	Twister	495900000	88000000	1996	PG-13	117	Action	Contemporary Fiction	57	6.0000

Figure 7. Chat2VIS analyzes data and shows results in visualization with a GPT prompt like "plot the gross against budget" [23].

The study in [75] examines the application of generative large language models (GLLMs) like ChatGPT and GPT-4 in accounting research. It emphasizes GLLMs' ability to perform complex textual analysis tasks, including those previously achievable only through human coding. The paper discusses the practicalities of using GLLMs in research, focusing on model selection, prompt engineering, construct validity, and addressing bias, replicability, and data privacy concerns. A case study is included to demonstrate GLLMs' capabilities, specifically in detecting non-answers in earnings conference calls, a task challenging for traditional automation.

The study in [76] focuses on the application of GPT models in data analysis, specifically in the context of discrete mathematics education. The study compares the performance of GPT-3.5 and GPT-4 in solving Proof Blocks problems, highlighting the significant improvement of GPT-4 over its predecessor. This comparison demonstrates the evolving ability of LLMs to handle complex academic and educational tasks, marking a notable advancement in the application of these models for intricate problem-solving and analysis in mathematics.

6. Research Design

GPT and LLMs can significantly assist researchers in research design for solving critical problems in various ways, as shown in [28,75,77–80]. They facilitate solving complex problems by providing insights into abstract reasoning tasks, enhancing the ability to conceptualize and tackle intricate issues. LLMs excel in deep textual analysis and detecting nuanced patterns, making them invaluable for research involving large volumes of text.

GPT models, especially the latest versions, show promise in solving advanced mathematical problems, aiding in disciplines that require rigorous analytical skills. They offer new

methodologies for conducting research, particularly in fields where traditional approaches are limited or inefficient.

In essence, GPT and LLMs open new avenues for addressing critical research challenges, offering tools that combine deep learning, language understanding, and problem-solving capabilities as summarized in Table 8.

Table 8. Review of existing literature in GPT-based research design.

Reference	Applicability of GPT/LLM Assists in Research	Example from Document
[77]	Explores GPT's capabilities in abstract reasoning and problem-solving with a focus on the Abstraction and Reasoning Corpus (ARC).	Investigates GPT's performance and challenges in solving simple abstract reasoning problems.
[28]	Uses LLMs to augment research on the P versus NP problem, proposing a Socratic reasoning framework for complex problem-solving with GPT-4.	Pilot study on P vs. NP problem showing GPT-4's capability in developing reasoning pathways.
[75]	Highlights the use of LLMs in complex textual analysis and problem-solving in various domains, including applications in finance and accounting.	Analyzes CEO humor in conference calls and clusters topics in employee reviews using LLMs.
[78]	Details the challenges and strategies of using GPT for solving ARC tasks, emphasizing the role of structured representations and external tools.	Demonstrates the enhancement of GPT's problem-solving abilities with object-based representations.
[79]	Examines ChatGPT's performance in solving verbal insight problems, comparing it to human problem-solving abilities in psychological research.	Evaluates ChatGPT's potential in solving complex problem-solving tasks.
[80]	Investigates GPT-3.5 and GPT-4's effectiveness in solving Proof Blocks problems in discrete mathematics, demonstrating GPT-4's significant improvement.	Examines GPT-4's success in mathematical proof solving, with a particular focus on combinatorics.

7. Results and Discussion

In accordance with the research questions delineated in Section 2, a sophisticated query mechanism was introduced in Figure 3. The implementation of this query method varies across databases used to obtain literature, as each database adheres to its own prescribed formulation for queries. This study utilized popular databases supporting advanced queries, including Scopus, IEEE Xplore, PubMed, Web of Science, and the ACM Digital Library, to compile a comprehensive list of literature. Table 9 details the advanced query applied to each database, along with the number of records retrieved on 30 January 2024, when these advanced queries were executed.

Notably, certain databases like Google Scholar were excluded due to the limitations of their advanced query mechanism. As elucidated in [81], Google Scholar lacks advanced search capabilities, the ability to download data, and faces challenges related to quality control and clear indexing guidelines. Consequently, Google Scholar is recommended for use as a supplementary source rather than a primary source for writing systematic literature reviews [82]. As shown in Table 9, the number of records retrieved from Scopus, IEEE Xplore, PubMed, Web of Science, and the ACM Library was 99, 119, 47, 306, and 102, respectively. Conversely, the use of the same advanced query in Google Scholar yielded over 30,000 results, underscoring the challenges associated with quality control and indexing guidelines within Google Scholar, as noted in [81,82].

In addition to utilizing the prominent databases outlined in Table 9, an advanced citation visualization tool, Litmaps, was employed to identify additional relevant studies [29]. The incorporation of LitMaps in the context of a systematic literature review is indispensable, given its pivotal role in augmenting the efficiency and efficacy of the review process. LitMaps functions as a sophisticated analytical tool facilitating the systematic organization, categorization, and visualization of an extensive corpus of scholarly literature, encompassing both citing and cited articles. As depicted in Figure 8, Litmaps identified

nine studies not present in Scopus, IEEE Xplore, PubMed, Web of Science, or the ACM Digital Library.

Table 9. Variation in advanced queries against each of the databases.

Database Source	Advanced Query	Records Returned
Scopus	TITLE-ABS-KEY ((gpt OR llm) AND research AND (design OR analyse OR (data AND augment*) OR (data AND generat*) OR feature) AND (problem OR gap OR question)) AND PUBYEAR > 2019 AND PUBYEAR < 2025 AND (LIMIT-TO (LANGUAGE, "English"))	99
IEEE Explore	((GPT OR LLM) AND Research AND (design OR analys* OR (data AND augment*) OR (data AND generat*) OR feature) AND (problem OR gap OR question))	119
PubMed	((GPT OR LLM) AND Research AND (design OR analys* OR (data AND augment*) OR (data AND generat*) OR feature) AND (problem OR gap OR question))	47
Web of Science	ALL = ((GPT OR LLM) AND Research AND (design OR analys* OR (data AND augment*) OR (data AND generat*) OR feature) AND (problem OR gap OR question))	306
ACM Digital Library	[[Abstract: gpt] OR [Abstract: llm]] AND [Abstract: research] AND [[Abstract: design] OR [Abstract: analys*] OR [Abstract: data] OR [Abstract: augment*] OR [[Abstract: data] AND [Abstract: generat*]] OR [Abstract: feature]] AND [[Abstract: problem] OR [Abstract: gap] OR [Abstract: question]] AND [E-Publication Date: (01/01/2019 to 31/12 2024)]	102

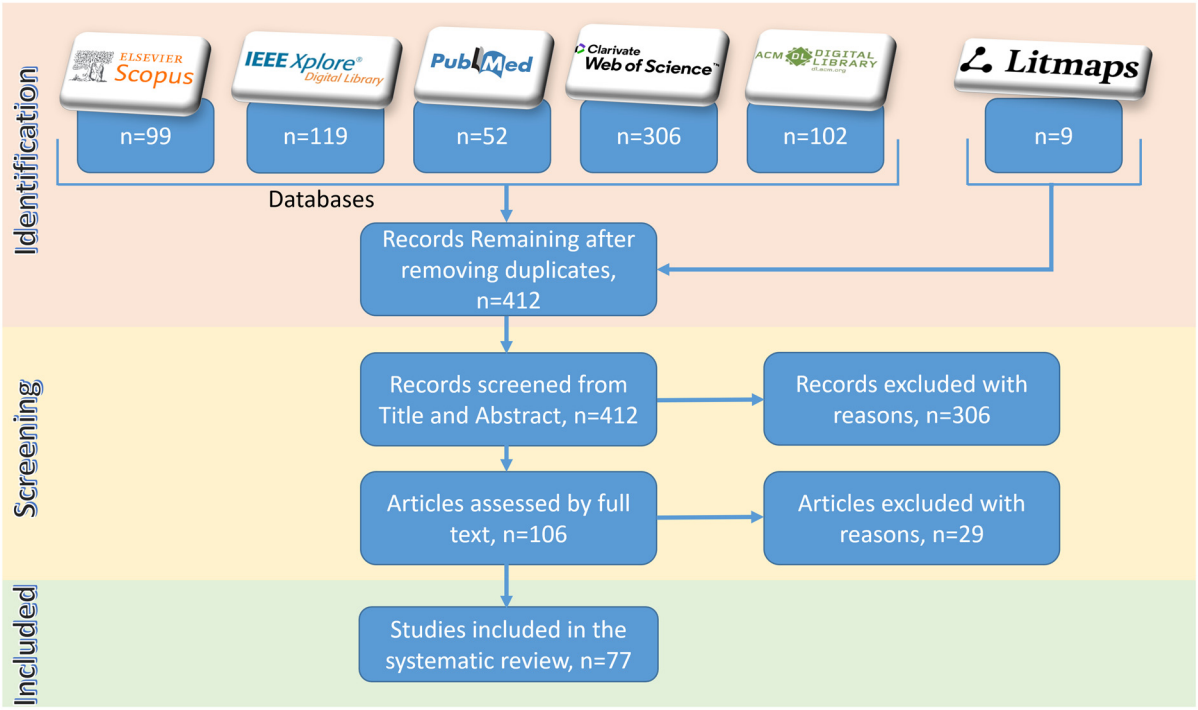

Figure 8. PRISMA Flow Diagram on the systematic literature review of "GPT for research".

The PRISMA flow diagram in Figure 8 illustrates the identification process, where 412 studies were initially identified after the deduplication of 275 studies. Following a screening process based on title and abstracts, 306 records were excluded due to their lack of relevance to the theme of "using GPT in performing research", despite the inclusion of keywords like GPT or research in the abstract or title. These studies were found to belong to entirely different areas, with some using the term "research" merely to denote future research endeavors or directions. Following the initial screening, full texts were obtained for 106 studies. Subsequently, a detailed analysis led to the exclusion of 29 full-text articles

as they did not address the original research questions regarding GPT's role in generating and processing research data, analyzing research data, or contributing to research design and problem-solving. Ultimately, 77 studies were included in the systematic literature review. While a few of these studies did not directly answer the original research questions, they provided valuable insights into the limitations, issues, and challenges associated with the adoption of GPT technologies in research activities.

This study conducted a thorough examination of over 77 peer-reviewed papers within the domain of "GPT in Research." Given the recent and trending nature of the GPT topic, a majority of the relevant papers span the past four years. Consequently, a significant 65% of the scrutinized papers were published in the years 2023 and 2024. The distribution includes 2 papers in 2024, 48 papers in 2023, 18 in 2022, 7 in 2021, and 4 in 2020, as illustrated in Figure 9.

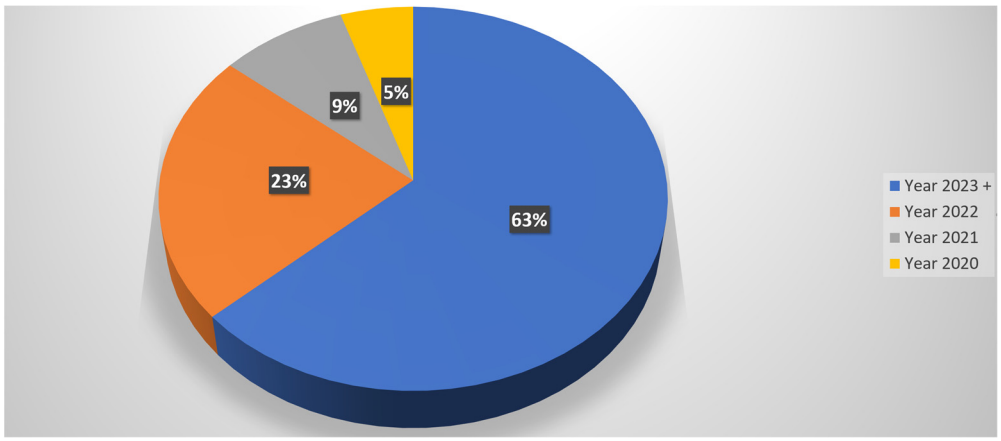

Figure 9. Timeline analysis of existing literature on the use of GPT in research.

To ensure the rigor and reliability of this systematic literature review, the following measures were meticulously implemented:

- Adherence to Established Guidelines: The review methodologically aligned with established guidelines and recommendations from seminal academic works, particularly the study in [82]. The work in [82], categorized Scopus, PubMed, Web of Science, ACM Digital Library, etc. as "Principal Sources". Hence, we exclusively used these databases.
- Exclusion of Unreliable Sources: A stringent criterion was applied to exclude sources lacking in quality control and those without clear indexing guidelines to maintain the overall reliability of the review. Existing scholarly work in [81,82] identified Google Scholar as one of the sources that lacks clear indexing guidelines. Hence, Google Scholar was not used as a source.
- Utilization of Advanced Visualization Tools: Advanced visualization tools, exemplified by Litmaps, played a pivotal role in assessing the alignment of identified studies with the predetermined domain of interest [29]. Litmaps facilitated a comprehensive evaluation, highlighting potential gaps and identifying relatively highly cited studies crucial for inclusion in the review process. Figure 8 illustrates the outcomes of utilizing Litmaps, showcasing the identification of nine significantly cited studies, thereby enhancing the comprehensiveness of the systematic literature review.
- Strategic Citation Analysis: In addition to advanced visualization tools, a strategic citation analysis was conducted to ascertain the prominence and impact of selected studies within the scholarly landscape. High-quality studies with a substantial number of citations were accorded due attention, contributing to the refinement and validation of the literature survey (using Litmaps).

8. Conclusions

Existing reviews on GPT [19–21] did not address how GPT could be useful for researchers in generating or augmenting research-related data and analyzing it. To mitigate this gap, this methodically crafted literature offers a strategic focus on data augmentation, backed by a meticulous examination of 412 scholarly works. In conclusion, the practical contributions of this comprehensive literature review are paramount in guiding researchers towards the judicious integration of GPT and associated technologies in their scholarly pursuits. By meticulously distilling 77 selected research contributions and developing a rigorous classification framework for "GPT's use on research data", this study provides a nuanced understanding of the multifaceted applications of GPT in data augmentation, critical analysis, and research design. Researchers can leverage the findings to inform their approach to generating and processing research data, analyzing complex datasets, and enhancing research design and problem-solving. Moreover, the systematic comparison of 54 extant literary works, evaluating diverse research domains, methodological approaches, and associated advantages and disadvantages, offers a practical roadmap for scientists seeking to seamlessly integrate GPT across various phases of their academic endeavors, thereby fostering innovation and efficiency in scholarly pursuits.

The deployment of GPT in research is not immune to inherent limitations, notably encompassing the issues of ethics [6], biases [7], hallucinations [83] and sycophantic behavior [73]. GPT, while proficient at generating human-like text, is susceptible to generating content that may be speculative or diverge from factual accuracy, leading to hallucinations within the generated information [83]. Furthermore, the model may exhibit sycophantic tendencies, showcasing an inclination to excessively praise or flatter, potentially compromising the objectivity and reliability of the generated output [73]. The manifestation of hallucinations and sycophantic behavior raises concerns about the model's capacity to maintain a rigorous and unbiased approach in generating content for research purposes, necessitating careful scrutiny and consideration of these limitations in the utilization of GPT within the academic realm.

Future studies could explore refining GPT through advanced training techniques to minimize bias, and hallucinations and enhance content accuracy. Additionally, research focusing on developing tailored algorithms to mitigate sycophantic behavior in GPT-generated content may contribute to more objective and reliable outputs for academic applications.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No dataset was generated in this research.

Acknowledgments: Special appreciation is reserved for the COEUS Institute, Maine, US (https://coeus.institute/ accessed on 27 January 2024), where the author serves as the Chief Technology Officer, for fostering an environment conducive to research excellence. The Institute's commitment to advancing knowledge in AI and pattern discovery has significantly enriched the author's endeavors in conducting this systematic review.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

GPT	Generative Pre-trained Transformer
LLM	Large Language Models
ML	Machine Learning
NER	Named Entity Recognition
NLP	Natural Language Processing
EHR	Electronic Health Records
GReaT	Generation of Realistic Tabular data

Appendix A

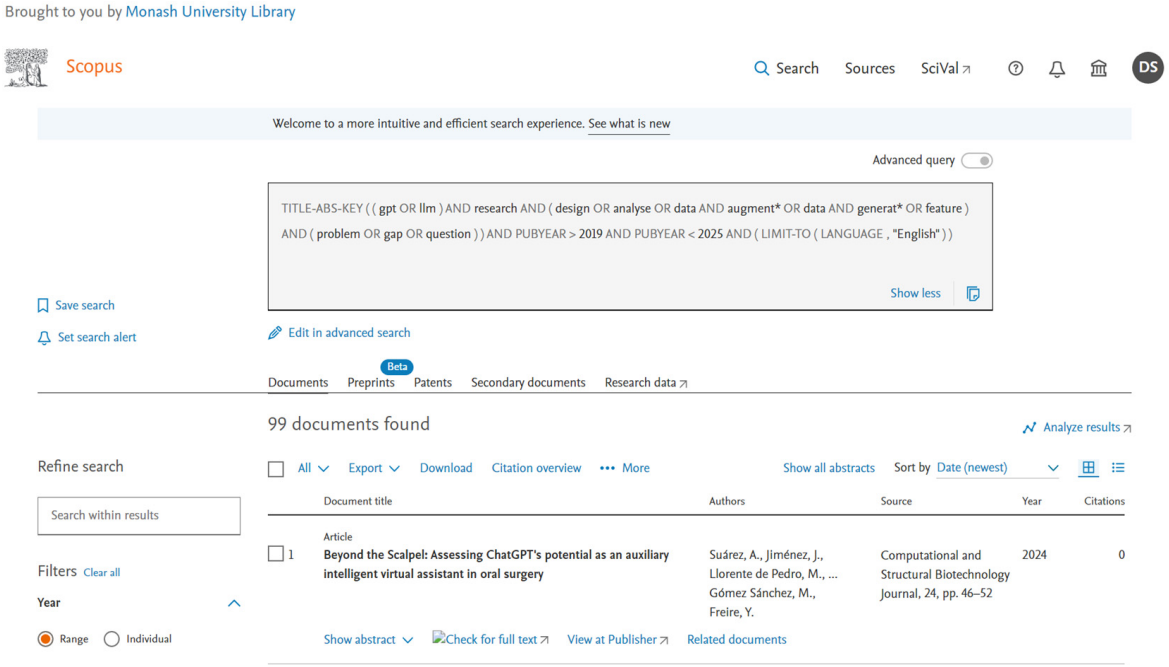

Figure A1. Database search from Scopus using Scopus-specific advanced query. From Scopus, 99 documents were returned, including the duplicates. After removing the duplicates, records were screened. For example, the first record, "Beyond the Scalpel: Assessing ChatGPT's Potential as an Auxiliary Intelligent Virtual Assistant in Oral Surgery" is not relevant to the focus of this study, "i.e., GPT in Research/GPT in Data Augmentation/GPT in Data Generation/GPT in Solving Research Problem".

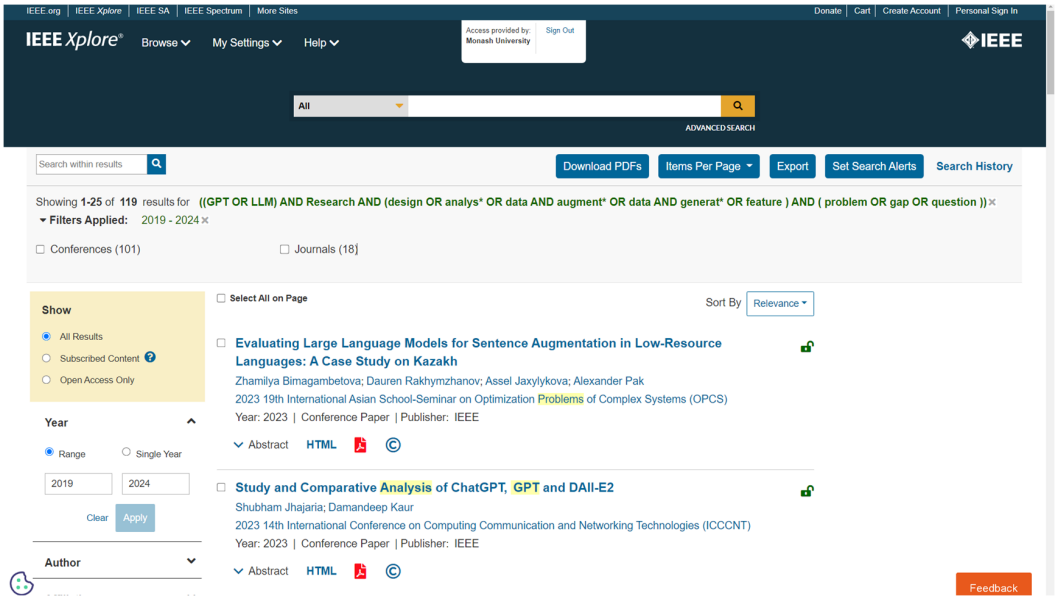

Figure A2. Database search from IEEE Xplore using IEEE Xplore-specific advanced queries. A total of 119 documents were returned, including duplicates. After removing the duplicates, records were screened. For example, the first record was included, and the second record was screened out as this paper does not address "GPT in research".

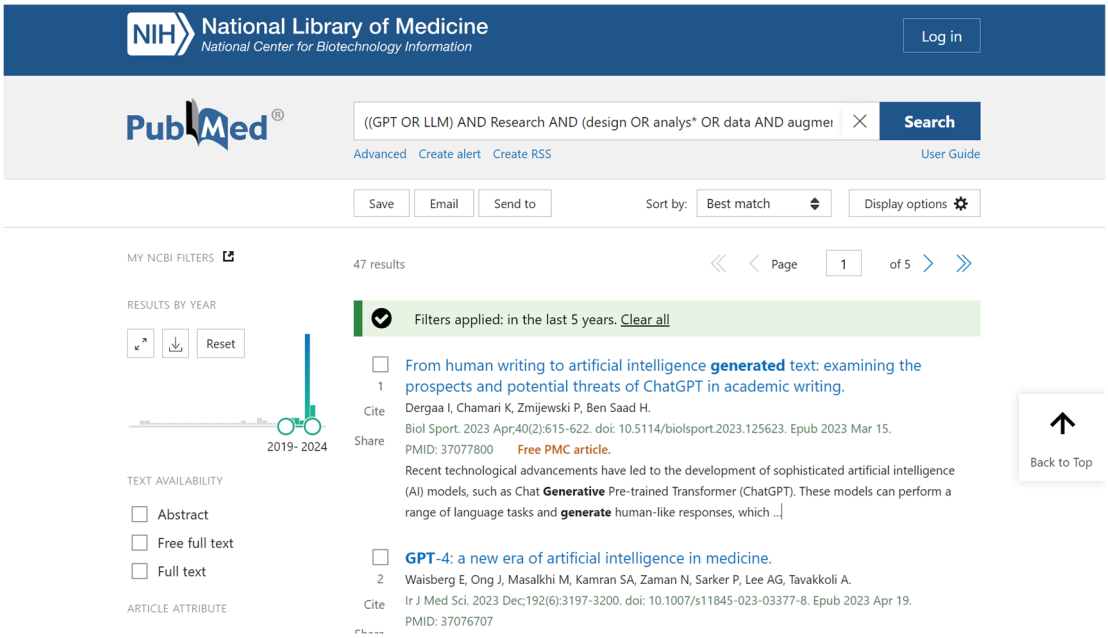

Figure A3. Database search from PubMed using a PubMed-specific advanced query. From PubMed, 47 documents were returned, including the duplicates.

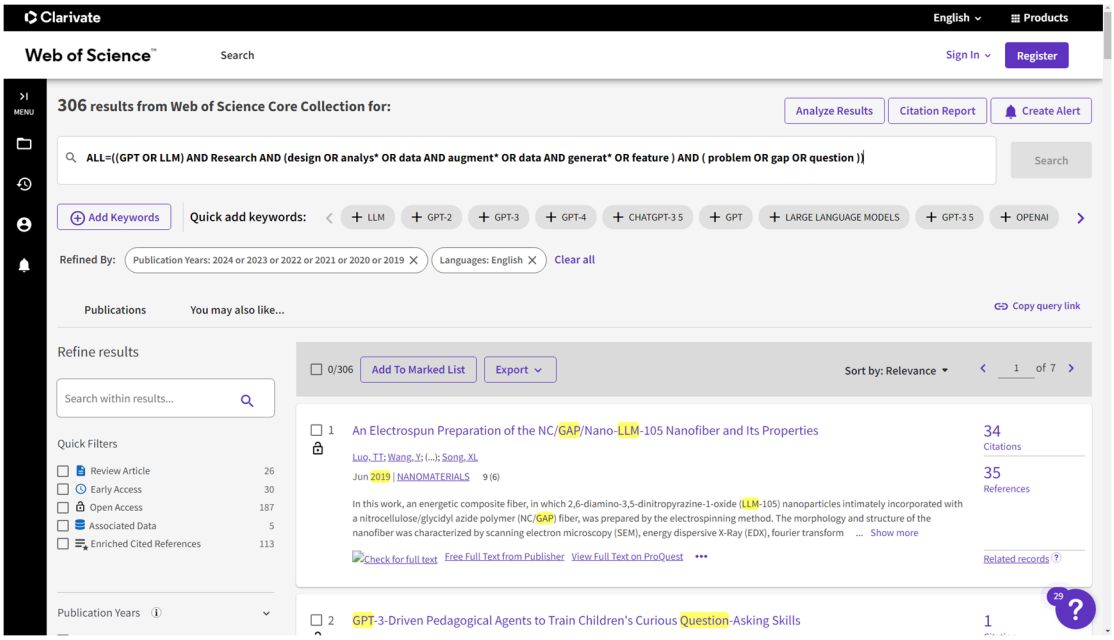

Figure A4. Database Search from Web of Science using their supported advanced query. From Web of Science, 306 documents were returned, including duplicates. After removing the duplicates, the records were screened. For example, the first records were screened out as this paper was focused on nanotechnology and nanomaterials.

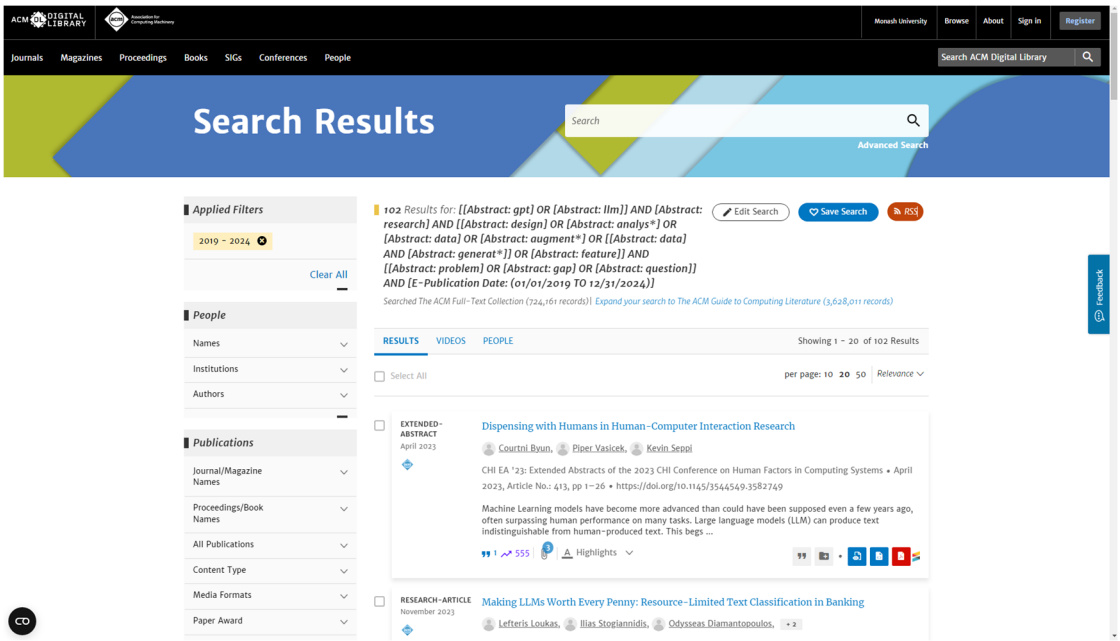

Figure A5. Database search from the ACM Digital Library using their supported advanced query. From the ACM Digital Library, 102 documents were returned, including duplicates.

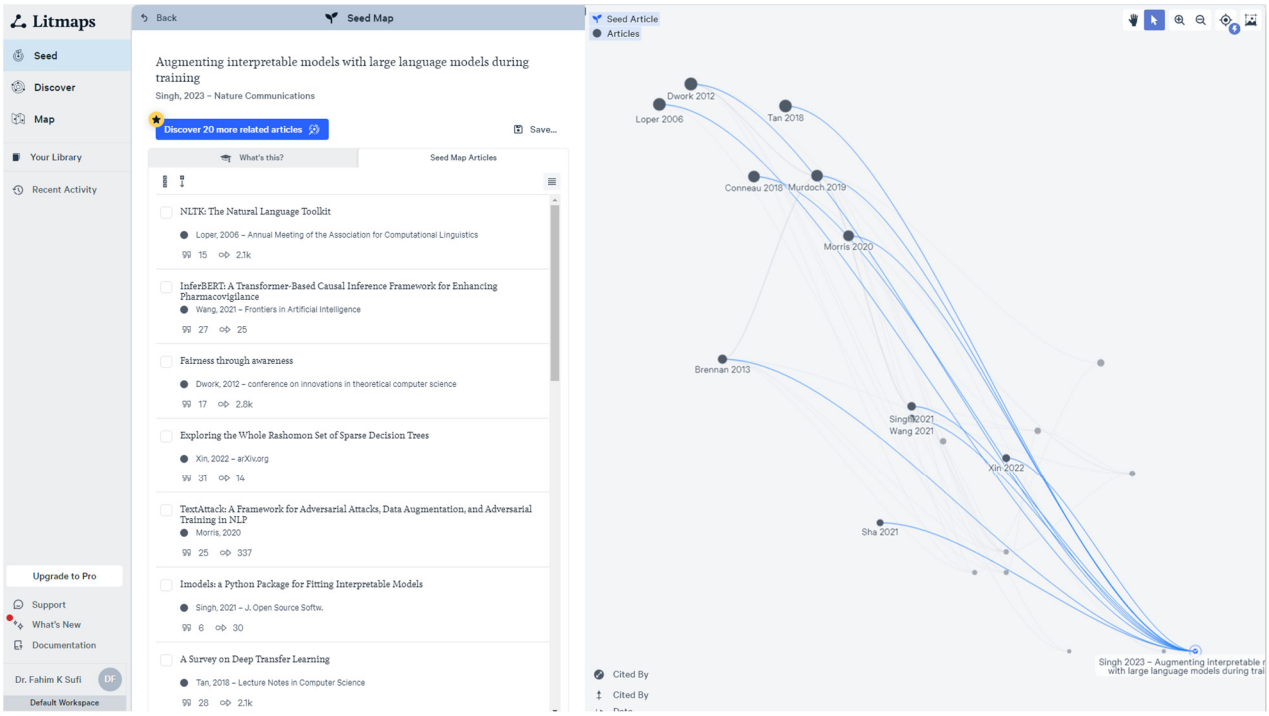

Figure A6. Litmaps suggest 20 possibly relevant articles by visually analyzing the citation maps of [66].

References

1. Adiguzel, T.; Kaya, M.H.; Cansu, F.K. Revolutionizing education with AI: Exploring the transformative potential of ChatGPT. *Contemp. Educ. Technol.* **2023**, *15*, ep429. [CrossRef]
2. Zhang, M.; Li, J. A commentary of GPT-3 in MIT Technology Review 2021. *Fundam. Res.* **2021**, *1*, 831–833. [CrossRef]
3. Katar, O.; Özkan, D.; Yildirim, Ö.; Acharya, U.R. Evaluation of GPT-3 AI Language Model in Research Paper Writing. *Turk. J. Sci. Technol.* **2023**, *18*, 311–318. [CrossRef]
4. Shibani, A.; Rajalakshmi, R.; Mattins, F.; Selvaraj, S.; Knight, S. Visual Representation of Co-Authorship with GPT-3: Studying Human-Machine Interaction for Effective Writing. In Proceedings of the 16th International Conference on Educational Data Mining, Bengaluru, India, 11–14 July 2023. [CrossRef]
5. Iorga, D. Journal of Comparative Research in Anthropology and Sociology Let Me Write That for You: Prospects Concerning the Impact of GPT-3 on the Copywriting Workforce. 2022. Available online: http://compaso.eu (accessed on 5 February 2024).
6. Watkins, R. Guidance for researchers and peer-reviewers on the ethical use of Large Language Models (LLMs) in scientific research workflows. *AI Ethics* **2023**. [CrossRef]
7. Casal, J.E.; Kessler, M. Can linguists distinguish between ChatGPT/AI and human writing?: A study of research ethics and academic publishing. *Res. Methods Appl. Linguist.* **2023**, *2*, 100068. [CrossRef]
8. Meyer, J.G.; Urbanowicz, R.J.; Martin, P.C.; O'Connor, K.; Li, R.; Peng, P.C.; Bright, T.J.; Tatonetti, N.; Won, K.J.; Gonzalez-Hernandez, G.; et al. ChatGPT and large language models in academia: Opportunities and challenges. *BioData Min.* **2023**, *16*, 20. [CrossRef] [PubMed]
9. Hosseini, M.; Resnik, D.B.; Holmes, K. The ethics of disclosing the use of artificial intelligence tools in writing scholarly manuscripts. *Res. Ethics* **2023**, *19*, 449–465. [CrossRef]
10. Pouran, A.; Veyseh, B.; Dernoncourt, F.; Min, B.; Nguyen, T.H. Generating Complement Data for Aspect Term Extraction with GPT-2. In Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing, Seattle, WA, USA, 14–15 July 2022.
11. Lu, Q.; Dou, D.; Nguyen, T.H. Textual Data Augmentation for Patient Outcomes Prediction. In Proceedings of the 2021 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2021, Houston, TX, USA, 9–12 December 2021; pp. 2817–2821. [CrossRef]
12. Kieser, F.; Wulff, P.; Kuhn, J.; Küchemann, S. Educational data augmentation in physics education research using ChatGPT. *Phys. Rev. Phys. Educ. Res.* **2023**, *19*, 020150. [CrossRef]
13. Sufi, F.K.; Alsulami, M.; Gutub, A. Automating Global Threat-Maps Generation via Advancements of News Sensors and AI. *Arab. J. Sci. Eng.* **2023**, *48*, 2455–2472. [CrossRef]
14. Sufi, F. Social Media Analytics on Russia–Ukraine Cyber War with Natural Language Processing: Perspectives and Challenges. *Information* **2023**, *14*, 485. [CrossRef]

15. Sufi, F.K.; Razzak, I.; Khalil, I. Tracking Anti-Vax Social Movement Using AI-Based Social Media Monitoring. *IEEE Trans. Technol. Soc.* **2022**, *3*, 290–299. [CrossRef]
16. Sufi, F.K.; Khalil, I. Automated Disaster Monitoring from Social Media Posts Using AI-Based Location Intelligence and Sentiment Analysis. *IEEE Trans. Comput. Soc. Syst.* **2022**. [CrossRef]
17. Sufi, F.K. AI-SocialDisaster: An AI-based software for identifying and analyzing natural disasters from social media. *Softw. Impacts* **2022**, *13*, 100319. [CrossRef]
18. Sufi, F. A decision support system for extracting artificial intelligence-driven insights from live twitter feeds on natural disasters. *Decis. Anal. J.* **2022**, *5*, 100130. [CrossRef]
19. Mahuli, S.A.; Rai, A.; Mahuli, A.V.; Kumar, A. Application ChatGPT in conducting systematic reviews and meta-analyses. *Br. Dent. J.* **2023**, *235*, 90–92. [CrossRef] [PubMed]
20. Sallam, M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. *Healthcare* **2023**, *11*, 887. [CrossRef] [PubMed]
21. Yenduri, G.; Ramalingam, M.; Chemmalar Selvi, G.; Supriya, Y.; Srivastava, G.; Maddikunta, P.K.; Deepti Raj, G.; Jhaveri, R.H.; Prabadevi, B.; Wang, W. Generative Pre-trained Transformer: A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions. *arxiv* **2023**, arXiv:2305.10435. [CrossRef]
22. Espejel, J.L.; Ettifouri, E.H.; Alassan, M.S.Y.; Chouham, E.M.; Dahhane, W. GPT-3.5, GPT-4, or BARD? Evaluating LLMs reasoning ability in zero-shot setting and performance boosting through prompts. *Nat. Lang. Process. J.* **2023**, *5*, 100032. [CrossRef]
23. Maddigan, P.; Susnjak, T. Chat2VIS: Generating Data Visualizations via Natural Language Using ChatGPT, Codex and GPT-3 Large Language Models. *IEEE Access* **2023**, *11*, 45181–45193. [CrossRef]
24. Sharma, A.; Devalia, D.; Almeida, W.; Patil, H.; Mishra, A. Statistical Data Analysis using GPT3: An Overview. In Proceedings of the 2022 IEEE Bombay Section Signature Conference (IBSSC), Mumbai, India, 8–10 December 2022.
25. Del, M.; Fishel, M. True Detective: A Deep Abductive Reasoning Benchmark Undoable for GPT-3 and Challenging for GPT-4. *arxiv* **2023**, arXiv:2212.10114. [CrossRef]
26. Jansen, B.J.; Jung, S.; Salminen, J. Employing large language models in survey research. *Nat. Lang. Process. J.* **2023**, *4*, 100020. [CrossRef]
27. Ai, M.R.; Quantum, M.A. The Impact of Large Language Models on Scientific Discovery: A Preliminary Study Using GPT-4. *arXiv* **2023**, arXiv:2311.07361.
28. Dong, Q.; Dong, L.; Xu, K.; Zhou, G.; Hao, Y.; Sui, Z.; Wei, F. Large Language Model for Science: A Study on P vs. NP. *arXiv* **2023**, arXiv:2309.05689.
29. Kaur, A.; Gulati, S.; Sharma, R.; Sinhababu, A.; Chakravarty, R. Visual citation navigation of open education resources using Litmaps. *Libr. Hi Tech. News* **2022**, *39*, 7–11. [CrossRef]
30. Sufi, F. Algorithms in Low-Code-No-Code for Research Applications: A Practical Review. *Algorithms* **2023**, *16*, 108. [CrossRef]
31. Borisov, V.; Seßler, K.; Leemann, T.; Pawelczyk, M.; Kasneci, G. Language Models are Realistic Tabular Data Generators. *arXiv* **2022**, arXiv:2210.06280.
32. Nakamoto, R.; Flanagan, B.; Yamauchi, T.; Dai, Y.; Takami, K.; Ogata, H. Enhancing Automated Scoring of Math Self-Explanation Quality Using LLM-Generated Datasets: A Semi-Supervised Approach. *Computers* **2023**, *12*, 217. [CrossRef]
33. Joon, J.; Chung, Y.; Kamar, E.; Amershi, S. Increasing Diversity While Maintaining Accuracy: Text Data Generation with Large Language Models and Human Interventions. *arxiv* **2023**, arXiv:2306.04140. [CrossRef]
34. Borisov, V.; Leemann, T.; Sessler, K.; Haug, J.; Pawelczyk, M.; Kasneci, G. Deep Neural Networks and Tabular Data: A Survey. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**. [CrossRef] [PubMed]
35. Acharya, A.; Singh, B.; Onoe, N. LLM Based Generation of Item-Description for Recommendation System. In Proceedings of the 17th ACM Conference on Recommender Systems, RecSys 2023, Singapore, 18–22 September 2023; pp. 1204–1207. [CrossRef]
36. Narayan, A.; Chami, I.; Orr, L.; Arora, S.; Ré, C. Can Foundation Models Wrangle Your Data? *arXiv* **2022**, arXiv:2205.09911. [CrossRef]
37. Bayer, M.; Kaufhold, M.A.; Buchhold, B.; Keller, M.; Dallmeyer, J.; Reuter, C. Data augmentation in natural language processing: A novel text generation approach for long and short text classifiers. *Int. J. Mach. Learn. Cybern.* **2023**, *14*, 135–150. [CrossRef]
38. Balaji, S.; Magar, R.; Jadhav, Y.; Farimani, A.B. GPT-MolBERTa: GPT Molecular Features Language Model for molecular property prediction. *arXiv* **2023**, arXiv:2310.03030.
39. Anaby-Tavor, A.; Carmeli, B.; Goldbraich, E.; Kantor, A.; Kour, G.; Shlomov, S.; Tepper, N.; Zwerdling, N. Do Not Have Enough Data? Deep Learning to the Rescue! *arxiv* **2019**, arXiv:1911.03118. [CrossRef]
40. Amin-Nejad, A.; Ive, J.; Velupillai, S. Exploring Transformer Text Generation for Medical Dataset Augmentation. In Proceedings of the Twelfth Language Resources and Evaluation Conference, Marseille, France, 19 May 2020; pp. 4699–4708. Available online: https://aclanthology.org/2020.lrec-1.578 (accessed on 5 February 2024).
41. Queiroz Abonizio, H.; Barbon Junior, S. Pre-trained Data Augmentation for Text Classification. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Springer Science and Business Media Deutschland GmbH: Cham, Switzerland, 2020; pp. 551–565. [CrossRef]
42. Cohen, S.; Presil, D.; Katz, O.; Arbili, O.; Messica, S.; Rokach, L. Enhancing social network hate detection using back translation and GPT-3 augmentations during training and test-time. *Inf. Fusion* **2023**, *99*, 101887. [CrossRef]

43. Chen, H.; Zhang, W.; Cheng, L.; Ye, H. Diverse and High-Quality Data Augmentation Using GPT for Named Entity Recognition. In *Communications in Computer and Information Science*; Springer Science and Business Media Deutschland GmbH: Singapore, 2023; pp. 272–283. [CrossRef]
44. Chang, Y.; Zhang, R.; Pu, J. I-WAS: A Data Augmentation Method with GPT-2 for Simile Detection. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Springer Science and Business Media Deutschland GmbH: Cham, Switzerland, 2023; pp. 265–279. [CrossRef]
45. Casula, C.; Tonelli, S.; Kessler, F.B. Generation-Based Data Augmentation for Offensive Language Detection: Is It Worth It? In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, Dubrovnik, Croatia, 2–6 May 2023; pp. 3359–3377. Available online: https://aclanthology.org/2023.eacl-main.244 (accessed on 5 February 2024).
46. Bird, J.J.; Pritchard, M.; Fratini, A.; Ekart, A.; Faria, D.R. Synthetic Biological Signals Machine-Generated by GPT-2 Improve the Classification of EEG and EMG through Data Augmentation. *IEEE Robot. Autom. Lett.* **2021**, *6*, 3498–3504. [CrossRef]
47. Hong, X.-S.; Wu, S.-H.; Tian, M.; Jiang, J. CYUT at the NTCIR-16 FinNum-3 Task: Data Resampling and Data Augmentation by Generation. In Proceedings of the 16th NTCIR Conference on Evaluation of Information Access Technologies, Tokyo Japan, 14–17 June 2022; pp. 95–102. Available online: http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings16/pdf/ntcir/03-NTCIR16-FINNUM-HongX.pdf (accessed on 5 February 2024).
48. Hassani, H.; Silva, E.S. The Role of ChatGPT in Data Science: How AI-Assisted Conversational Interfaces Are Revolutionizing the Field. *Big Data Cogn. Comput.* **2023**, *7*, 62. [CrossRef]
49. Grasler, I.; Preus, D.; Brandt, L.; Mohr, M. Efficient Extraction of Technical Requirements Applying Data Augmentation. In Proceedings of the ISSE 2022—2022 8th IEEE International Symposium on Systems Engineering, Vienna, Austria, 24–26 October 2022. [CrossRef]
50. D'Sa, A.G.; Illina, I.; Fohr, D.; Klakow, D.; Ruiter, D. Exploring Conditional Language Model Based Data Augmentation Approaches for Hate Speech Classification. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Springer Science and Business Media Deutschland GmbH: Cham, Switzerland, 2021; pp. 135–146. [CrossRef]
51. Hu, Y.; Mai, G.; Cundy, C.; Choi, K.; Lao, N.; Liu, W.; Lakhanpal, G.; Zhou, R.Z.; Joseph, K. Geo-knowledge-guided GPT models improve the extraction of location descriptions from disaster-related social media messages. *Int. J. Geogr. Inf. Sci.* **2023**, *37*, 2289–2318. [CrossRef]
52. Khatri, S.; Iqbal, M.; Ubakanma, G.; Van Der Vliet-Firth, S. SkillBot: Towards Data Augmentation using Transformer language model and linguistic evaluation. In Proceedings of the 2022 International Conference on Human-Centered Cognitive Systems, HCCS 2022, Shanghai, China, 17–18 December 2022. [CrossRef]
53. Elbadawi, M.; Li, H.; Basit, A.W.; Gaisford, S. The role of artificial intelligence in generating original scientific research. *Int. J. Pharm.* **2024**, *652*, 123741. [CrossRef] [PubMed]
54. Maharana, A.; Bansal, M. GRADA: Graph Generative Data Augmentation for Commonsense Reasoning. In Proceedings of the 29th International Conference on Computational Linguistics, Gyeongju, Republic of Korea, 12–17 October 2022.
55. Hämäläinen, P.; Tavast, M.; Kunnari, A. Evaluating Large Language Models in Generating Synthetic HCI Research Data: A Case Study. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, Hamburg, Germany, 23–28 April 2023; ACM: New York, NY, USA, 2023; pp. 1–19. [CrossRef]
56. Maimaiti, M.; Liu, Y.; Luan, H.; Sun, M. Data augmentation for low-resource languages NMT guided by constrained sampling. *Int. J. Intell. Syst.* **2022**, *37*, 30–51. [CrossRef]
57. Meyer, S.; Elsweiler, D.; Ludwig, B.; Fernandez-Pichel, M.; Losada, D.E. Do We Still Need Human Assessors' Prompt-Based GPT-3 User Simulation in Conversational AI. In *ACM International Conference Proceeding Series*; Association for Computing Machinery: New York, NY, USA, 2022. [CrossRef]
58. Modzelewski, A.; Sosnowski, W.; Wilczynska, M.; Wierzbicki, A. DSHacker at SemEval-2023 Task 3: Genres and Persuasion Techniques Detection with Multilingual Data Augmentation through Machine Translation and Text Generation. In Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), Toronto, Canada, 15 July 2023; pp. 1582–1591. Available online: https://aclanthology.org/2023.semeval-1.218 (accessed on 5 February 2024).
59. Nouri, N. Data Augmentation with Dual Training for Offensive Span Detection. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Seattle, WA, USA, 10–15 July 2022.
60. Pellicer, L.F.A.O.; Ferreira, T.M.; Costa, A.H.R. Data augmentation techniques in natural language processing. *Appl. Soft Comput.* **2023**, *132*, 109803. [CrossRef]
61. Van Nooten, J.; Daelemans, W. Improving Dutch Vaccine Hesitancy Monitoring via Multi-Label Data Augmentation with GPT-3.5. In Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, Toronto, Canada, July 2023; pp. 251–270. Available online: https://aclanthology.org/2023.wassa-1.23 (accessed on 5 February 2024).
62. Romero-Sandoval, M.; Calderón-Ramírez, S.; Solís, M. Using GPT-3 as a Text Data Augmentator for a Complex Text Detector. In Proceedings of the 2023 IEEE 5th International Conference on BioInspired Processing (BIP), San Carlos, Alajuela, Costa Rica, 28–30 November 2023; pp. 1–6. [CrossRef]

63. Rebboud, Y.; Lisena, P.; Troncy, R. Prompt-based Data Augmentation for Semantically-Precise Event Relation Classification. In Proceedings of the SEMMES 2023, Semantic Methods for Events and Stories, Heraklion, Greece, 23–28 May 2023; Available online: https://www.eurecom.fr/publication/7298 (accessed on 5 February 2024).
64. Quteineh, H.; Samothrakis, S.; Sutcliffe, R. Textual Data Augmentation for Efficient Active Learning on Tiny Datasets. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 19–20 November 2020; pp. 7400–7410. Available online: https://aclanthology.org/2020.emnlp-main.600 (accessed on 5 February 2024).
65. Suhaeni, C.; Yong, H.S. Mitigating Class Imbalance in Sentiment Analysis through GPT-3-Generated Synthetic Sentences. *Appl. Sci.* **2023**, *13*, 9766. [CrossRef]
66. CSingh; Askari, A.; Caruana, R.; Gao, J. Augmenting interpretable models with large language models during training. *Nat. Commun.* **2023**, *14*, 7913. [CrossRef]
67. Veyseh, A.P.B.; Van Nguyen, M.; Min, B.; Nguyen, T.H. Augmenting Open-Domain Event Detection with Synthetic Data from GPT-2. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Springer Science and Business Media Deutschland GmbH: Cham, Switzerland, 2021; pp. 644–660. [CrossRef]
68. Tapia-Téllez, J.M.; Escalante, H.J. Data Augmentation with Transformers for Text Classification. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Springer Science and Business Media Deutschland GmbH: Cham, Switzerland, 2020; pp. 247–259. [CrossRef]
69. Zhou, S.; Zhang, Y. DATLMedQA: A data augmentation and transfer learning based solution for medical question answering. *Appl. Sci.* **2021**, *11*, 1251. [CrossRef]
70. Waisberg, E.; Ong, J.; Kamran, S.A.; Masalkhi, M.; Zaman, N.; Sarker, P.; Lee, A.G.; Tavakkoli, A. Bridging artificial intelligence in medicine with generative pre-trained transformer (GPT) technology. *J. Med. Artif. Intell.* **2023**, *6*. [CrossRef]
71. Vogel, L.; Flek, L. Investigating Paraphrasing-Based Data Augmentation for Task-Oriented Dialogue Systems. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Springer Science and Business Media Deutschland GmbH: Cham, Switzerland, 2022; pp. 476–488. [CrossRef]
72. Sawai, R.; Paik, I.; Kuwana, A. Sentence augmentation for language translation using gpt-2. *Electronics* **2021**, *10*, 3082. [CrossRef]
73. Ranaldi, L.; Pucci, G. When Large Language Models contradict humans? Large Language Models' Sycophantic Behaviour. *arXiv* **2023**, arXiv:2311.09410.
74. Lim, S.; Schmälzle, R. Artificial intelligence for health message generation: An empirical study using a large language model (LLM) and prompt engineering. *Front. Commun.* **2023**, *8*, 1129082. [CrossRef]
75. de Kok, T. Generative LLMs and Textual Analysis in Accounting: (Chat)GPT as Research Assistant? *SSRN* **2023**. [CrossRef]
76. Lengerich, B.J.; Bordt, S.; Nori, H.; Nunnally, M.E.; Aphinyanaphongs, Y.; Kellis, M.; Caruana, R. LLMs Understand Glass-Box Models, Discover Surprises, and Suggest Repairs. *arXiv* **2023**, arXiv:2308.01157.
77. Arora, D.; Singh, H.G. Have LLMs Advanced Enough? A Challenging Problem Solving Benchmark for Large Language Models Mausam IIT Delhi. *arxiv* **2023**, arXiv:2305.15074. [CrossRef]
78. Xu, Y.; Li, W.; Vaezipoor, P.; Sanner, S.; Khalil, E.B. LLMs and the Abstraction and Reasoning Corpus: Successes, Failures, and the Importance of Object-based Representations. *arXiv* **2023**, arXiv:2305.18354.
79. Orrù, G.; Piarulli, A.; Conversano, C.; Gemignani, A. Human-like problem-solving abilities in large language models using ChatGPT. *Front. Artif. Intell.* **2023**, *6*, 1199350. [CrossRef] [PubMed]
80. Poulsen, S.; Sarsa, S.; Prather, J.; Leinonen, J.; Becker, B.A.; Hellas, A.; Denny, P.; Reeves, B.N. Solving Proof Block Problems Using Large Language Models. In Proceedings of the SIGCSE 2024, Portland, OR, USA, 20–23 March 2024; Volume 7.
81. Halevi, G.; Moed, H.; Bar-Ilan, J. Suitability of Google Scholar as a source of scientific information and as a source of data for scientific evaluation—Review of the Literature. *J. Informetr.* **2017**, *11*, 823–834. [CrossRef]
82. Gusenbauer, M.; Haddaway, N.R. Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of Google Scholar, PubMed, and 26 other resources. *Res. Synth. Methods* **2020**, *11*, 181–217. [CrossRef] [PubMed]
83. Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.J.; Madotto, A.; Fung, P. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* **2023**, *55*, 1–38. [CrossRef]