future internet

MDPI

Review

A Systematic Review and Comprehensive Analysis of Pioneering AI Chatbot Models from Education to Healthcare: ChatGPT, Bard, Llama, Ernie and Grok

Ketmanto Wangsa [1], Shakir Karim [2], Ergun Gide [2] and Mahmoud Elkhodr [2,*]

[1] Independent Researcher, Sydney 2000, Australia; wangsaketmanto@gmail.com
[2] School of Engineering and Technology, Central Queensland University, Rockhampton 4701, Australia; s.karim@cqu.edu.au (S.K.); e.gide@cqu.edu.au (E.G.)
* Correspondence: m.elkhodr@cqu.edu.au

Abstract: AI chatbots have emerged as powerful tools for providing text-based solutions to a wide range of everyday challenges. Selecting the appropriate chatbot is crucial for optimising outcomes. This paper presents a comprehensive comparative analysis of five leading chatbots: ChatGPT, Bard, Llama, Ernie, and Grok. The analysis is based on a systematic review of 28 scholarly articles. The review indicates that ChatGPT, developed by OpenAI, excels in educational, medical, humanities, and writing applications but struggles with real-time data accuracy and lacks open-source flexibility. Bard, powered by Google, leverages real-time internet data for problem solving and shows potential in competitive quiz environments, albeit with performance variability and inconsistencies in responses. Llama, an open-source model from Meta, demonstrates significant promise in medical contexts, natural language processing, and personalised educational tools, yet it requires substantial computational resources. Ernie, developed by Baidu, specialises in Chinese language tasks, thus providing localised advantages that may not extend globally due to restrictive policies. Grok, developed by Xai and still in its early stages, shows promise in providing engaging, real-time interactions, humour, and mathematical reasoning capabilities, but its full potential remains to be evaluated through further development and empirical testing. The findings underscore the context-dependent utility of each model and the absence of a singularly superior chatbot. Future research should expand to include a wider range of fields, explore practical applications, and address concerns related to data privacy, ethics, security, and the responsible deployment of these technologies.

Keywords: AI chatbots; ChatGPT; Bard; Llama; Ernie; Grok; systematic review; comparative analysis; natural language processing; real-time data; computational resources; educational applications; medical applications; language specialisation

check for updates

Citation: Wangsa, K.; Karim, S.; Gide, E.; Elkhodr, M. A Systematic Review and Comprehensive Analysis of Pioneering AI Chatbot Models from Education to Healthcare: ChatGPT, Bard, Llama, Ernie and Grok. *Future Internet* **2024**, *16*, 219. https://doi.org/10.3390/fi16070219

Academic Editor: Ivan Serina

Received: 23 May 2024
Revised: 13 June 2024
Accepted: 20 June 2024
Published: 22 June 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The Large Language Model (LLM) has revolutionised conversational AI by employing a two-step process: pre-training on extensive human-curated data followed by fine-tuning with human instructions. This approach enables the generation of diverse, human-like text. Chatbots represent the principal application of LLMs. OpenAI's development of ChatGPT, powered by the Generative Pre-Trained Transformer (GPT) model, marked a pivotal moment in this revolution in September 2022 [1]. As a generative AI language model, ChatGPT synthesises and arranges words based on user inputs, thus pushing the boundaries of human–computer interaction [2].

Following the introduction of ChatGPT, other major technology firms have entered the fray, thus receiving public endorsement [3,4]. Google introduced Bard, which is powered by the LaMDA and PaLM models [5,6]. Other significant contenders include Baidu's Ernie, Facebook's Llama, and Xai's Grok. Figure 1 depicts the development timeline of these language models and chatbots.

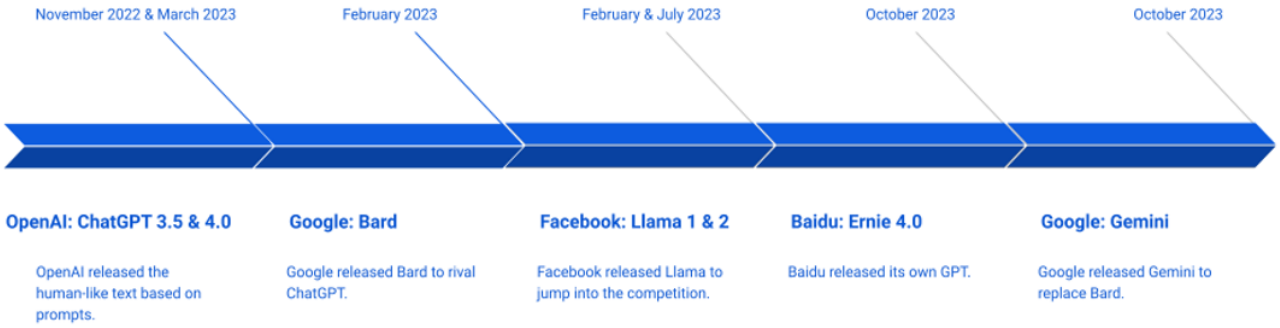

Figure 1. Timeline of language model and chatbot development.

To date, no studies have provided a comprehensive overview comparing ChatGPT, Bard, Llama, Ernie, and Grok within a single paper. For clarity, this paper refers to ChatGPT 3.5 and ChatGPT 4.0 collectively as "ChatGPT", and similar amalgamations are applied to other chatbots. Specific versions, such as ChatGPT 3.5 or Bard's different iterations, are discussed in Section 4. Bing Chat is excluded from this study due to its reliance on the same underlying technology, GPT 4, as ChatGPT 4. The term "Llama" is uniformly used to prevent confusion. Although prior studies have compared subsets of these chatbots, none have encompassed all five in an unified analysis. Agarwal et al. [7] compared ChatGPT, Bard, and Bing in the context of medical physiology but did not include Ernie, Llama, or Grok. Other studies have similarly focused on limited aspects or subsets of these technologies [8–11]. This study aims to fill the gap in the literature by providing a detailed comparative analysis of ChatGPT, Bard, Llama, Ernie, and Grok, thus exploring their applications across various fields. Consequently, this study aims to answer the followings research questions:

1. How do the unique features, target audiences, and potential future developments of ChatGPT, Bard, Llama, Ernie, and Grok compare, and what insights can be drawn from this analysis to guide the selection and application of AI chatbots in various domains?
2. What are the key ethical concerns, cybersecurity risks, and limitations associated with the use of AI chatbots, and how can these challenges be addressed to ensure the responsible development and deployment of these technologies?
3. Considering the differentiated strengths and weaknesses of each AI chatbot across various domains, such as medicine, technology, education, ethics, humanities, and writing, what are the implications for the context-dependent utility of these models, and how can this knowledge inform future research and practice in the field of AI language models?

By addressing these research questions, this study aims to provide a comprehensive review and analysis of the current state and future potential of AI chatbots, thus contributing to the ongoing scholarly dialogue and informing the responsible development and application of these technologies in diverse global contexts. Hence, the unique contributions of this study include the following:

1. Conducting a comprehensive comparative analysis of five leading AI chatbots—ChatGPT, Bard, Llama, Ernie, and Grok—across multiple domains, including medicine, technology, education, ethics, humanities, and general human interaction, based on a systematic review of 28 scholarly articles using evaluation criteria such as performance metrics, scope of knowledge, computational efficiency, and domain expertise rating.
2. Identifying and synthesising the specific strengths, weaknesses, and contextual applications of each chatbot based on the findings from the reviewed literature, thus providing a nuanced and evidence-based understanding of their utility and limitations in various practical scenarios.

3. Highlighting the rapid evolution of AI technologies and the need for continuous, updated evaluations while addressing critical issues such as data privacy, security, potential biases, inconsistencies in responses, and the need for clear guidelines and regulations to ensure the ethical deployment of AI language models.

2. Research Methodology

A systematic literature review was employed as the primary methodology for sourcing the relevant literature, thus serving both as a procedural framework and a narrative for conducting the research [12]. This common approach involved searching for information, extracting it, analysing it, and subsequently reporting on the findings [13]. Modifications to this approach were noted in other studies [14–16]. Additionally, contributions from both online and printed literature reviews assisted in shaping the study's contributions. Utilising this literature, the paper presents a comparative analysis aimed at identifying the conceptual contents of the research field. However, the research papers in Section 4 were exclusively sourced from repositories that met predefined criteria.

Figure 2 illustrates the study's methodology. The research commenced by identifying gaps in the scientific community, notably the lack of comparative analyses of the chatbots: ChatGPT, Bard, Llama, Ernie, and Grok. Searches were conducted on two repositories, Semantic Scholar and ResearchGate, and included Grok's official website due to its emerging status and the absence of comprehensive papers discussing its strengths and weaknesses. Specific procedures were implemented to search the relevant literature by reviewing existing studies and providing a complete analysis of these studies [12]. The search keyword used was ("ChatGPT" AND "Bard" AND "Llama" AND "Ernie"). This process resulted in papers for four chatbots and one official website for one chatbot.

Upon retrieval, the papers were extracted to identify those that would contribute to and enrich the study. Each paper was reviewed, classified into specific fields, and assessed for the presence of the chatbots under study. The selected papers were then presented, and the findings from each were formulated, thus highlighting the strengths and weaknesses of the chatbots. These findings will be detailed in Section 4 and divided into two subsections: Findings and Discussion and Limitations, thus revealing the advantages and disadvantages of the selected chatbots. Thus, the final output is the comparative analysis of ChatGPT, Bard, Llama, Ernie, and Grok.

2.1. Literature Search Criteria

The literature papers were collected from the online repositories Semantic Scholar, which is an AI-powered research tool, and ResearchGate, which hosts over 160 million publications. The search criteria were developed following the guidelines for conducting a systematic literature review, as outlined by Kitchenham and Charters [17]. The keyword for the search on both platforms was ("ChatGPT" AND "Bard" AND "Llama" AND "Ernie"). For Grok, by Xai, as no public research has yet outlined Grok as a rising contender in the chatbot landscape, the focus was solely on its official site.

The terminology was standardised, such as referring to Google Bard as Gemini, and including the latest versions of each chatbot, such as ChatGPT 3.5 and 4.0, as well as Llama 13B, which is associated with Llama 7B and Llama 2. The metadata, including the title and abstract, were also considered. Publications were restricted to those dated within the last five years (2019–2024) to ensure the richness and accuracy of the papers. Duplicates were strictly excluded. The details are provided in Table 1.

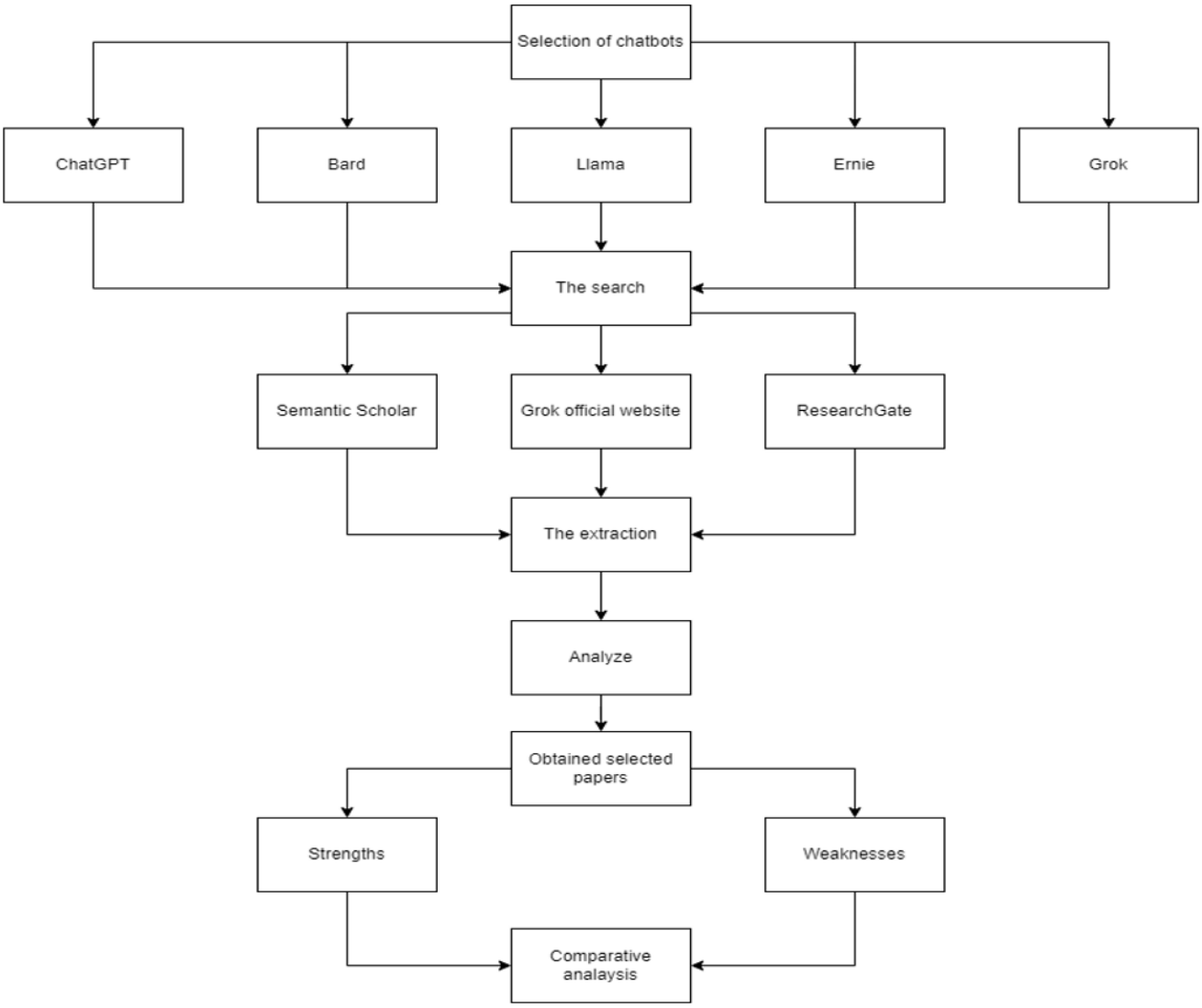

Figure 2. The methodological process of the study.

Table 1. Search criteria.

No	Criteria	Description
1	Database	Semantic Scholar and ResearchGate
2	Keyword of search	(“ChatGPT” AND “Bard” AND “Llama” AND “Ernie”)
3	Date of collection	As of 25 February 2024
4	Year of publication	In the last 5 years (2019–2024)
5	Type of publication	Peer-reviewed articles in a journal
6	Integrity	No duplicates
7	Title and abstract (focus)	Studies focus on the following criteria: ChatGPT, Bard, Llama, Ernie
8	Language	English only
9	Full-text analysis	Support our focus to create a comparative analysis between four language models

Based on Figure 3, the extraction process consisted of three stages: Identification, Screening, and Inclusion. Initially, 128 papers were identified across both platforms. During the Identification stage, 68 non-journal articles were eliminated, thus leaving 60 papers. The Screening stage further removed two duplicates and 26 papers with titles and abstracts that did not significantly contribute to the study, thus resulting in 32 papers. During the Screening stage, specific quality criteria were systematically applied. The titles of the papers were required to explicitly reference the name of the chatbot(s) or include relevant

keywords such as "AI", "generative AI", "AI chatbot", or "chatbot". The abstracts needed to clearly articulate either the benefits or the drawbacks of the chatbot(s) in question. During the full-text analysis phase, the papers were selected based on their contributions to understanding the advantages, disadvantages, comparisons, reviews, benchmarks, or implications regarding the performance of various chatbots across different fields.

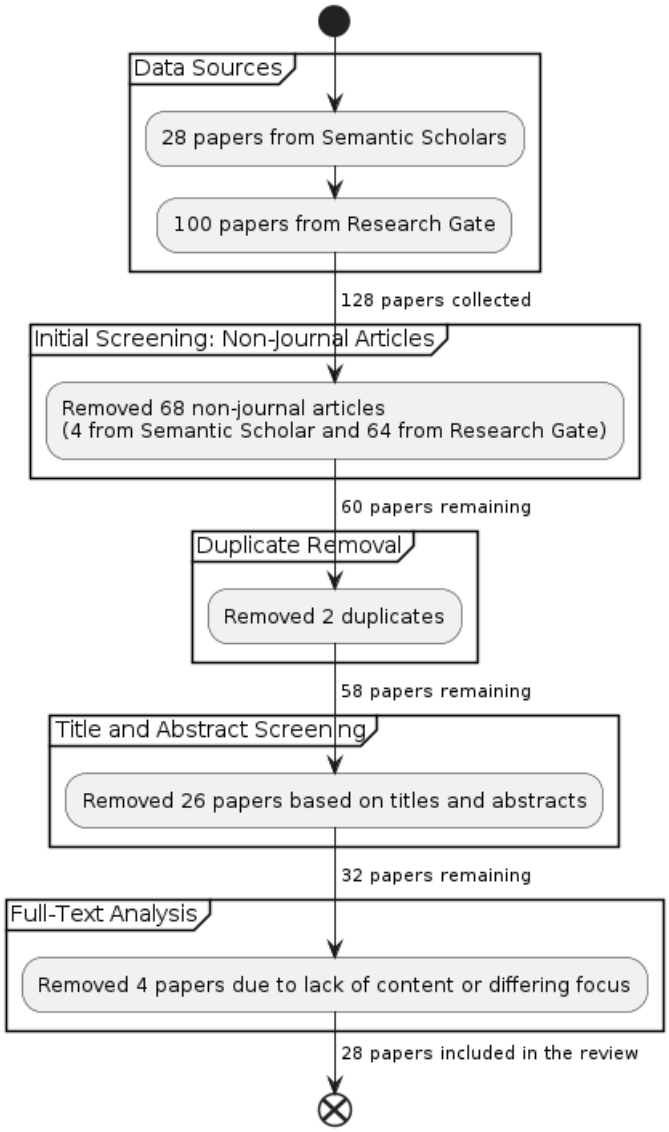

Figure 3. Stages of the extraction process.

ResearchGate was utilised as one of the sources for retrieving papers due to its comprehensive database of contemporary research, thus complementing the more traditional databases and providing access to recent studies that may not yet be indexed elsewhere. This approach ensured a comprehensive and up-to-date collection of the relevant literature.

2.2. The Extraction of Systematic Review Papers

Figure 4 illustrates the diverse fields contributing to our study. The research integrates six disciplines: medical terms such as physiology, dentistry, radiology, diseases, and medicine; technology terms, including applications, coding, computers, and artificial intelligence; academic writing; morality; and humanity. As depicted in Figure 4, the medical sector dominates with 13 papers, followed by technology with five contributions, thus underscoring significant engagement within these fields. Numerous scholars have explored

ChatGPT's application in these areas [18–23]. Education and morality are tied for third place with three papers each, while humanity and writing are fourth with two papers each.

Figure 5 categorises the language models and chatbots evaluated in this study. Based on a full-text analysis, five groups were identified. Of the 28 papers reviewed, 24 referenced ChatGPT, 21 referenced Bard, 7 referenced Llama, 2 referenced Ernie, and 10 referenced other language models and chatbots such as Bing, Watson, Claude, and Aria, which were excluded from this study.

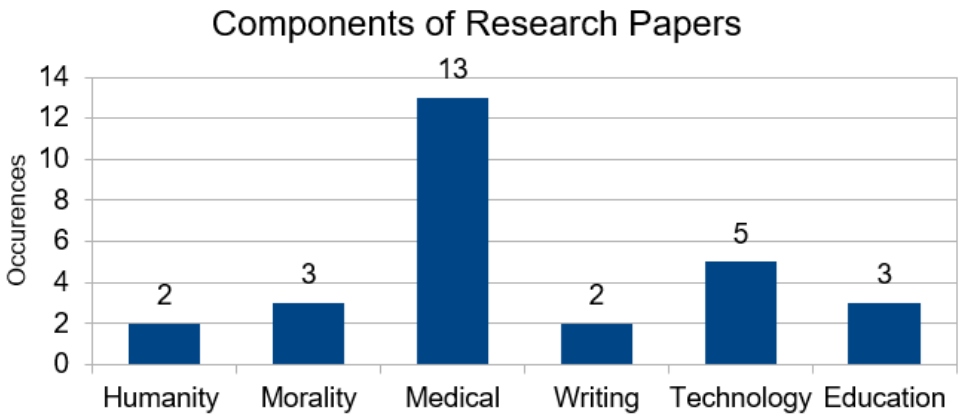

Figure 4. Components of systematic review research papers.

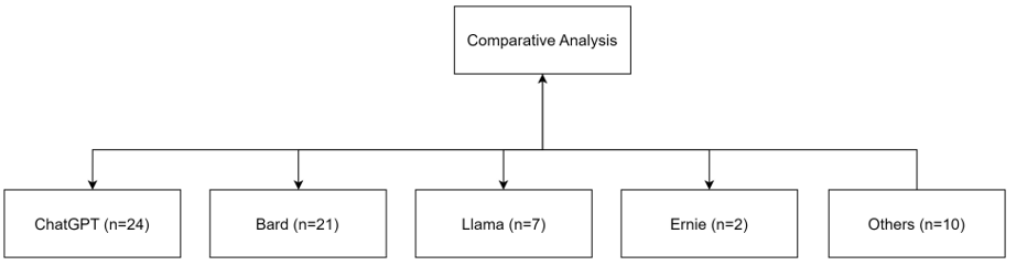

Figure 5. Composition of language models and chatbots.

3. Systematic Review

This section presents the theoretical background of the chatbots explored in this paper. The information was sourced from a variety of resources, including articles, news, and essential web pages. The aim is to establish a foundational understanding before presenting the findings in Section 4. It is important to note that due to the rapid evolution of AI chatbots, the details discussed may haved changed over time. For instance, the parameters of the chatbots might vary due to their limited public access at the time of writing. Therefore, this systematic review is introduced in good faith that it will remain relevant to this study.

3.1. Detailed Comparative Overview of Chatbot Models

Table 2 provides a comparative overview of the chatbot models discussed in this study, thus highlighting their publishers, base models, and parameter ranges. This information facilitates a deeper understanding of the technological breadth and specific capabilities of each model.

Table 2. Comparative overview of chatbot models.

Chatbots	ChatGPT	Bard	Ernie	Llama	Grok
Publisher	OpenAI	Google	Baidu	Facebook	Xai
Base Model	GPT 3.5, GPT 4.0	LaMDa, PaLM	Ernie 3.0, 4.0	Llama 1.0, 2.0	Grok 0, 1.0, 1.5
Parameters	Up to 1.76 T	520 B	260 B	65 B to 70 B	33 B to 314 B

Following the summary provided in Table 2, detailed explanations are offered for each chatbot model, thus enhancing the comprehension of their technological advancements and operational scopes.

3.1.1. OpenAI's ChatGPT

OpenAI developed ChatGPT based on the Generative Pre-Training Transformer (GPT) model. Notable models include ChatGPT 3.5 and ChatGPT 4.0, where ChatGPT 3.5 has 175 billion parameters, and ChatGPT 4.0, based on GPT 4.0, has a significantly larger number of parameters [19,24]. The ChatGPT series demonstrates OpenAI's progression in AI sophistication, thus featuring parameter counts escalating from 175 billion to as high as 1.76 trillion across various specialised versions. This extensive parameter range is indicative of the model's adaptability and its capability to manage diverse, intricate tasks effectively. Numerous studies have highlighted ChatGPT's positive outcomes and contributions, thus making it the most studied chatbot [9]. Despite its promising contributions, ChatGPT is not without controversy; for example, Italy banned ChatGPT due to privacy concerns [25], though the issues were later addressed by OpenAI to comply with regulatory policies. Other countries such as North Korea, Iran, Russia, and China have also imposed bans on ChatGPT.

3.1.2. Google's Bard

Google developed Bard to compete with ChatGPT, thus initially using the LaMDa model and later transitioning to PaLM, which boasts 520 billion parameters [26]. Bard, starting with the LaMDa model and later integrating the PaLM architecture, exemplifies Google's commitment to developing AI capable of sophisticated data interpretation and multifaceted response generation, thereby making it suitable for a broad array of content generation tasks. Since its launch on 21 March 2023, Bard has become available in over 230 countries, thus processing both open-ended and close-ended queries, generating creative content, and assisting in job-related tasks like drafting resumes [27]. Bard's unique integration with Google services, such as YouTube and Drive, allows it to provide up-to-date information directly from the internet.

3.1.3. Facebook's Llama

Meta introduced the Llama chatbot series, ranging from Llama 1.0 with 65 billion parameters to Llama 2.0, which offers models sized between 7 billion and 70 billion parameters [28]. Llama, with its variants, is designed to fulfill a variety of computational demands. This range allows for flexible deployment across different technological environments, despite lacking direct internet access. Llama 2.0 marks a significant advancement, which was developed in collaboration with Microsoft. Unlike Bard, Llama lacks direct internet access, and its performance is dependent on substantial computational resources [20].

3.1.4. Baidu's Ernie

Baidu launched Ernie as its flagship AI, thus offering several models as open-source on GitHub [29]. Specifically catering to the Chinese market, Ernie's models like Ernie 3.0 and the less-documented Ernie 4.0, with parameters numbering up to 260 billion, reflect Baidu's focus on integrating regional linguistic and cultural nuances into its AI solutions. Ernie is particularly tailored to the Chinese market, thus reflecting local cultural norms and language. While Ernie 3.0 is well documented with 260 billion parameters, the specifics of Ernie 4.0 remain undisclosed [4].

3.1.5. Xai's Grok

Elon Musk's Xai developed Grok in response to the perceived shortcomings in existing chatbots. Initiated by Elon Musk, Grok began with 33 billion parameters in Grok 0 and expanded to 314 billion in later versions such as Grok 1. This rapid development underlines Grok's capacity for high-level reasoning and real-time information processing, thus

significantly enhancing its utility in dynamic, data-driven applications. Grok 0 debuted with 33 billion parameters and has shown proficiency in mathematical reasoning and humour [30]. The Grok 1 model, introduced with 314 billion parameters, promises further enhancements and real-time connectivity with social media platforms [31].

4. Comparative Analysis of AI Language Models Based on Scholarly Literature

This section presents the evaluation criteria and methodology employed, followed by a detailed discussion of the key findings and implications derived from the comparative analysis of ChatGPT, Bard, Llama, Ernie, and Grok.

4.1. Evaluation Criteria and Methodology

Table 3 presents a structured analysis based on the synthesis of existing research, which highlights the distinctive strengths, opportunities, and limitations of the AI models ChatGPT, Bard, Llama, Ernie, and Grok. This synthesis is grounded in the following selected criteria: Strengths and Opportunities, Weaknesses and Challenges, Performance Metrics, Scope of Knowledge, Use Cases Excelling In, Use Cases Limited By, Knowledge Update Frequency, Computational Efficiency, and Domain Expertise Rating.

The categories of Strengths and Opportunities, Weaknesses and Challenges, Scope of Knowledge, and Use Cases have been directly extracted from an extensive array of academic papers analysed in this study. These dimensions represent the empirical findings and scholarly assessments articulated within the published literature. The judgements on Performance Metrics are not derived from our empirical experimentation but instead reflect a subjective synthesis based on the performance and behaviours of the models as reported in the literature. Similarly, the Scope of Knowledge is an interpretive assessment of the range of topics and the depth of discourse that each model demonstrates across various domains, as depicted in the scholarly works reviewed.

The categories of Use Cases Excelling In and Use Cases Limited By encapsulate instances of proficiency and deficiency as reported in the papers reviewed within their respective applications and evaluations. Knowledge Update Frequency has been estimated from the update and versioning information presented within these studies, which is a critical factor for the models' ongoing accuracy and relevance. Computational Efficiency represents a qualitative aggregate of how the models balance their functional prowess with the computational overhead required, which is crucial for assessing their real-world applicability. Lastly, the Domain Expertise Rating has been deduced from the literature's consensus on each model's command over specialised knowledge areas.

This evaluative framework, entirely based on a scholarly review of published research, does not include any new experimental data generated by the authors of this paper. It aims to offer an academic perspective rather than definitive quantitative results. It is an attempt to synthesise the reported observations and findings to compare and contrast the current capabilities and future potential of these prominent AI language models. Large-scale and longitudinal studies, as well as cross-disciplinary experimentation, are needed to scientifically validate these results. We hope these insights are viewed as part of the ongoing effort in this direction rather than a conclusive comparison between LLMs.

Table 3. Comparative analysis overview of chatbot models.

Criteria	ChatGPT	Bard	Llama	Ernie	Grok
Strengths and Opportunities	Multiple choice questions, Jeopardy!, academic text, patient education materials; tends to be an absolutist; superior in academic and practical tasks	Non-original problems online, interactive feedback, Jeopardy!, recipe correction, anatomy questions, "situationist", real-time internet access	Medical chatbot, visual content adjustment, NLP potential, predicts diagnosis groups, arithmetic problem solving, open source	Reads various Chinese languages, Baidu integration	Real-time information, humour, suggests questions, excels in math and reasoning, has 314 billion parameters, open source
Weaknesses and Challenges	Inaccurate information, consistency issues, academic concerns, lack of contextual awareness, safety concerns, drug interaction, original scientific contributions	Generates different answers, consistency issues, email consistency, anatomy writing, original scientific contributions	Not suitable for calculus and statistics, limited access to present time, requires more resources	Main focus on mainland China, strict Chinese policy may affect quality, limited access	Early development, exclusive to X premium subscribers, inaccurate information (bias)
Performance Metrics	High	Medium	Medium	High	Medium
Scope of Knowledge	High	High (Real-time)	Medium	Low	Low
Use Cases Excelling In	Academic research, trivia, education	Education, recipe correction, real-time queries	Healthcare, education, programming	Chinese language tasks, Baidu-related queries	Social media analysis, humour, math problems
Use Cases Limited By	Scientific contributions, drug interactions	Detailed scientific writing, complex reasoning tasks	Calculus, statistics, real-time data	International use, diverse datasets	General knowledge, extensive reasoning
Knowledge Update Frequency	High (Quarterly updates)	High (Real-time)	Medium (Yearly updates)	Low (Biennial updates)	Medium (Regular updates)
Computational Efficiency	Medium	High	Low	High	Medium
Domain Expertise Rating	High	Medium	Medium	High (in Chinese domains)	Low

4.2. Insights from ChatGPT's Bifurcation

The bifurcation of ChatGPT into 'general' and 'versions 3.5 and 4.0' classifications provides a more nuanced understanding of its capabilities, thus highlighting the advancements made in the later versions. This differentiation allows for a clearer comparison between the overarching strengths of ChatGPT and the specific enhancements in versions 3.5 and 4.0.

ChatGPT's proficiency in generating medical physiology multiple choice questions and evaluating educational materials for conditions like sleep apnoea underscores its potential as an educational tool, particularly for novice learners and those in the medical domain. The comparative analysis of ChatGPT and Bard's performance in the context of the Jeopardy! quiz show offers an engaging perspective on their knowledge retrieval and question-answering abilities, thus suggesting their potential as preparatory tools for participants.

ChatGPT's tendency towards absolutism in ethical positions contrasts with Bard's "situationist" philosophy, thus highlighting the differences in their approaches to ethical considerations and adaptability. The specific capabilities of ChatGPT 3.5 and 4.0 in mathematics, such as tackling original logic problems and addressing calculus and statistics challenges, emphasise their utility for students and researchers in problem solving and academic writing.

ChatGPT 3.5's proficiency in programming across multiple languages (Java, Python, and C++) and its ability to cater to both non-programmers and experienced coders further demonstrate its versatility and potential impact on coding education and practice. The medical domain expertise of ChatGPT 4.0, exemplified by its success in the Psychiatric Licensing Examination 2022 in Taiwan and its contributions to radiology, nephrology, neurosurgery, and ophthalmology, underscore its transformative potential in healthcare education and practice.

4.3. Llama and Ernie: Specialised Applications

Llama's potential to develop a ChatDoctor, predict diagnosis-related groups, and integrate with natural language processing models like BERT highlights its promising applications in medical knowledge dissemination and patient care. Ernie's proficiency in comprehending various Chinese dialects and its integration with Baidu position it as a powerful tool for serving the linguistic diversity of China's population and becoming an integral part of the country's digital ecosystem.

4.4. Grok: Emerging Potential

Grok's nascent but promising capabilities in real-time social media insights, engagement with "spicy" questions, and incorporation of humor set it apart from other LLMs, thus suggesting its potential for unique applications in social media analysis and user interaction.

4.5. Key Findings and Implications

The domain of language models exhibits a diverse array of capabilities, with each model possessing unique strengths that make them suitable for particular tasks. ChatGPT stands out with its high performance metrics and comprehensive scope of knowledge, thus excelling in academic environments. Its quarterly knowledge updates ensure its advice remains pertinent, especially in academic research, trivia, and education. However, it is challenged by issues of accuracy in areas involving drug interactions and scientific contributions.

Bard, with its real-time access to information and high computational efficiency, shines in providing current and interactive educational content. Its ability to assist with non-original problems online and correct recipes reflects a strong orientation towards practical, everyday tasks. Yet, Bard's medium domain expertise rating indicates room for growth, particularly in handling complex reasoning tasks and detailed scientific writing.

Llama's medium performance metrics reflect a balance between its capabilities as a medical chatbot and its need for more computational resources. It shows considerable promise in healthcare and education, though its medium domain expertise rating and limited suitability for advanced mathematical topics like calculus suggest targeted application within its scope.

Ernie exhibits high domain expertise in Chinese language tasks, with its high performance metrics indicating robustness within its specialised scope. However, its low scope of knowledge and biennial knowledge updates suggest its best use is within China-centric applications that do not require frequent updates or an international scope.

Lastly, Grok's playful interaction with social media content and humour indicates a model designed for engaging user experiences. Its medium computational efficiency and low domain expertise rating suggest that it is in an evolutionary phase, thus poised for growth as it receives regular updates. Grok's ability to suggest questions and handle humour creatively positions it as a versatile tool in social scenarios, though it is presently limited by its general knowledge and reasoning capabilities.

Across all models, the knowledge update frequency is a pivotal factor, thereby directly affecting their ability to provide timely and relevant responses. The computational efficiency also varies, thus influencing the practicality of their use in different environments. The domain expertise rating serves as a benchmark for each model's proficiency in specialised areas, which is crucial for tasks that demand in-depth knowledge.

The varied strengths and limitations of these AI models underscore the importance of selecting the right tool for the task at hand. ChatGPT's robust academic support contrasts with Bard's real-time interactivity, Llama's medical specialisation, Ernie's localised Chinese expertise, and Grok's emerging versatility. As the AI field advances, it is these distinctive features that will guide users in choosing the most appropriate model for their needs, be it for academic support, practical tasks, healthcare assistance, language-specific applications, or engaging social interactions. Table 4 provides a complementary perspective to the comparative analysis by highlighting the unique features, target audiences, and potential future developments of each AI language model. It offers insights into the distinctive characteristics that set these models apart, the specific user groups who may benefit the most from their capabilities, and the possible directions for their evolution and growth in the near future. The unique features column in Table 4 sheds light on the specific attributes that make each model stand out, such as ChatGPT's adaptive learning, Bard's contextual awareness, Ernie's multi-dialect processing, Llama's medical knowledge integration, and Grok's sentiment analysis capabilities. These features underscore the specialised capabilities and potential niche applications of each model.

Furthermore, the target audiences column in Table 4 helps identify the user groups who may find each model particularly relevant and beneficial. This information can guide decisionmakers, researchers, and users in selecting the most appropriate model for their specific needs and contexts. For instance, students and researchers may gravitate towards ChatGPT for its academic prowess, while healthcare professionals and patients may find Llama's medical knowledge integration more pertinent to their requirements.

Lastly, the potential future developments column in Table 4 explores these AI language models' possible trajectories and advancements. By considering the future potential, such as ChatGPT's integration with educational platforms, Bard's enhanced natural language understanding, Ernie's expansion to other Asian languages, Llama's collaboration with medical institutions, and Grok's real-time data analysis and visualisation capabilities, a forward-looking perspective on the evolving landscape of AI language models and their potential impact across various domains is provided.

Table 4. Unique features, target audiences, and potential future developments.

Chatbots	Unique Features	Target Audiences	Potential Future Developments
ChatGPT	Multiple choice questions, Jeopardy!, academic text, patient education materials; tends to be an absolutist; superior in academic and practical tasks	Academic researchers, students, educators	Integration with educational platforms, enhanced natural language understanding
	Inaccurate information, consistency issues, academic concerns, lack of contextual awareness, safety concerns, drug interaction, original scientific contributions		
Bard	Non-original problems online, interactive feedback, Jeopardy!, recipe correction, anatomy questions, "situationist", real-time internet access Generates different answers, consistency issues, email consistency, anatomy writing, original scientific contributions	General public, educators, students	Improved contextual awareness, real-time data processing
Llama	Medical chatbot, visual content adjustment, NLP potential, predicts diagnosis groups, arithmetic problem solving, open source	Healthcare professionals, educators, programmers	Collaboration with medical institutions, advanced NLP capabilities
	Not suitable for calculus and statistics, limited access to present time, requires more resources		
Ernie	Reads various Chinese languages, Baidu integration	Chinese speakers, local businesses, language learners	Expansion to other Asian languages, integration with international platforms
	Main focus on mainland China, strict Chinese policy may affect quality, limited access		
Grok	Real-time information, humor, suggests questions, excels in math and reasoning, has 314 billion parameters, open source	Social media users, general public, educators	Enhanced real-time data analysis, improved sentiment analysis
	Early development, exclusive to X premium subscribers, inaccurate information (bias)		

4.6. Comparative Analysis of AI Chatbots Using Mindmap

Figure 6 provides a mindmap that summarises the Comparative Analysis of AI Chatbots collected from the 28 articles included in this study. The mindmap, which was developed using Whimsical [32], provides a structured summary of these studies in evaluating the performance, ethical considerations, and applications of the AI chatbots involved in this study.

This comparative analysis highlights several key insights and findings across different domains as follows:

Ethical and Security Concerns: One of the primary concerns addressed is the ethical implications and cybersecurity risks associated with the use of AI chatbots. Over-reliance and over-trust in these systems can lead to significant security vulnerabilities. Implementing robust ethical practices is suggested to mitigate these risks.

Performance in Various Fields: The performance of AI chatbots varies across different fields. For example, in the Jeopardy Challenge, ChatGPT and Bard provided different answers to the same questions, thus demonstrating variability in their responses. In medical physiology, ChatGPT generated less difficult multiple choice questions compared to Bard. ChatGPT also excelled in providing coherent scientific writing but raised concerns about the accuracy and integrity of the generated text.

Specific AI Models: This analysis included evaluations of specific AI models such as ChatDoctor (based on the Llama 7B model) and DRG-LLaMA. ChatDoctor demonstrated

superior knowledge in medical terms, while DRG-LLaMA excelled in predicting diagnosis-related groups for hospitalised patients.

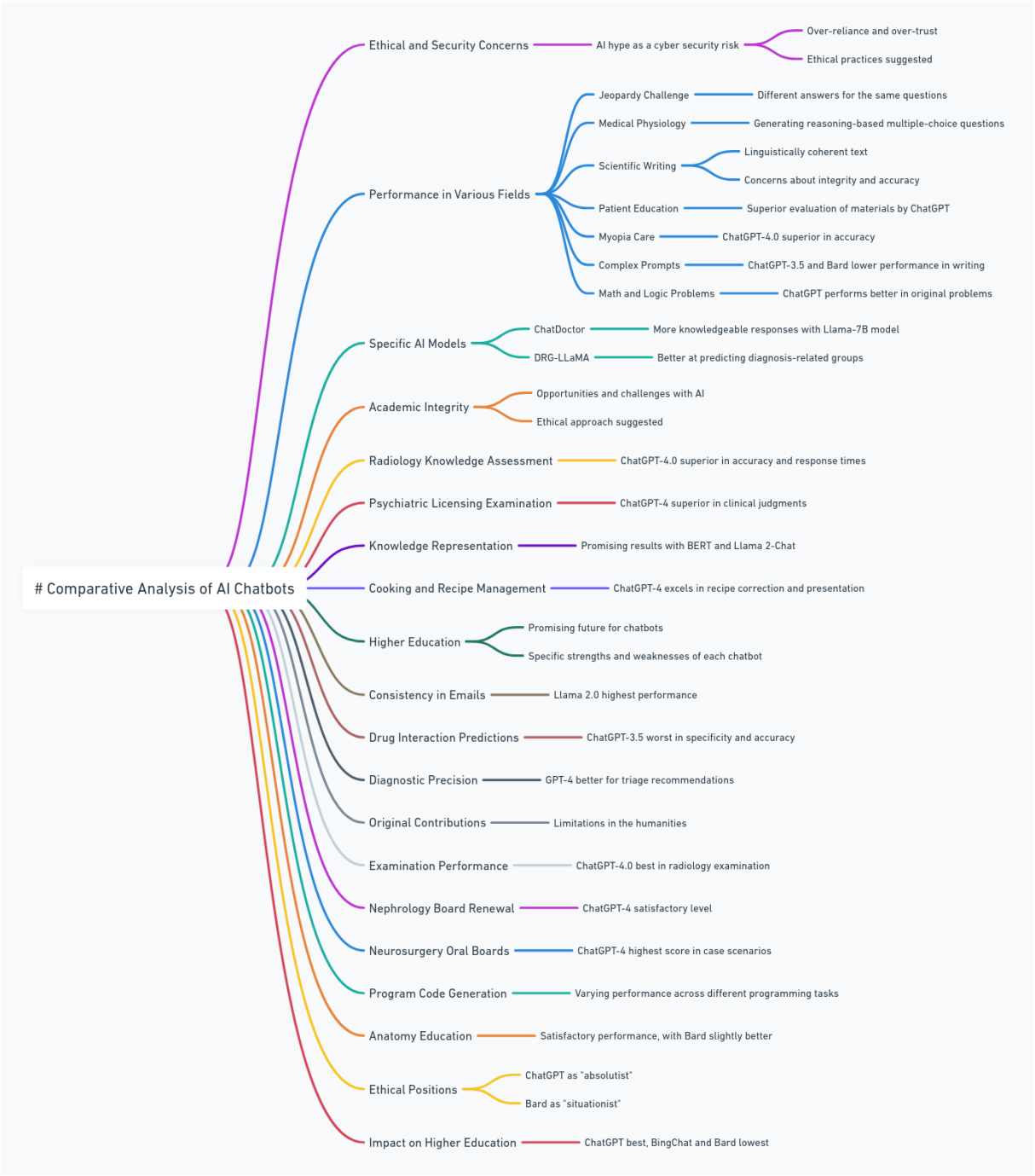

Figure 6. Composition of language models and chatbots Mindmap.

Academic Integrity: AI chatbots present both opportunities and challenges for academic integrity. They can generate various content types, thus potentially leading to issues like plagiarism. An ethical approach is essential to address these challenges effectively.

Examination Performance: ChatGPT 4.0 consistently outperformed other models in radiology and psychiatric licensing examinations, thus demonstrating superior accuracy and response times. It also achieved high scores in neurosurgery oral board preparations and nephrology board renewal standards.

Knowledge Representation and Natural Language Processing: Models like BERT and Llama 2 Chat showed promising results in enhancing natural language processing capabilities, thus furthering their applications in knowledge representation tasks.

Cooking and Recipe Management: In practical applications such as cooking, ChatGPT 4 excelled in recipe correction, time management, and presentation, thus showcasing its utility beyond traditional academic and clinical settings.

Impact on Higher Education: The overall impact of AI chatbots on higher education was found to be significant, with ChatGPT leading in performance, while models like BingChat and Bard showed relatively lower effectiveness.

These insights from the mindmap illustrate the diverse applications and varying performance levels of AI chatbots, thus emphasizing the need for ethical considerations and highlighting areas where these technologies excel or require improvement.

4.7. Heatmap of AI Chatbot Performance

4.7.1. Methodology Applied for Performance Score

Scoring was conducted by analysing the 28 papers involved in this study and by examining the Strengths and Weaknesses identified in the previous sections. Each criterion was rated on a scale from 1 to 10, with 10 representing the highest possible score and 1 the lowest. The criteria used for scoring are as follows:

- **Accuracy**: The correctness of responses generated by the chatbot.
- **Consistency**: The reliability and uniformity of responses across different instances.
- **Domain Expertise**: The chatbot's depth of knowledge in specialised areas such as medicine, engineering, etc.
- **Computational Efficiency**: How efficiently the chatbot uses computational resources.
- **Scope of Knowledge**: The breadth of topics and information the chatbot can accurately cover.

The heatmap (Figure 7) provides a visual representation of these scores, thus highlighting areas of strength and weakness for each chatbot.

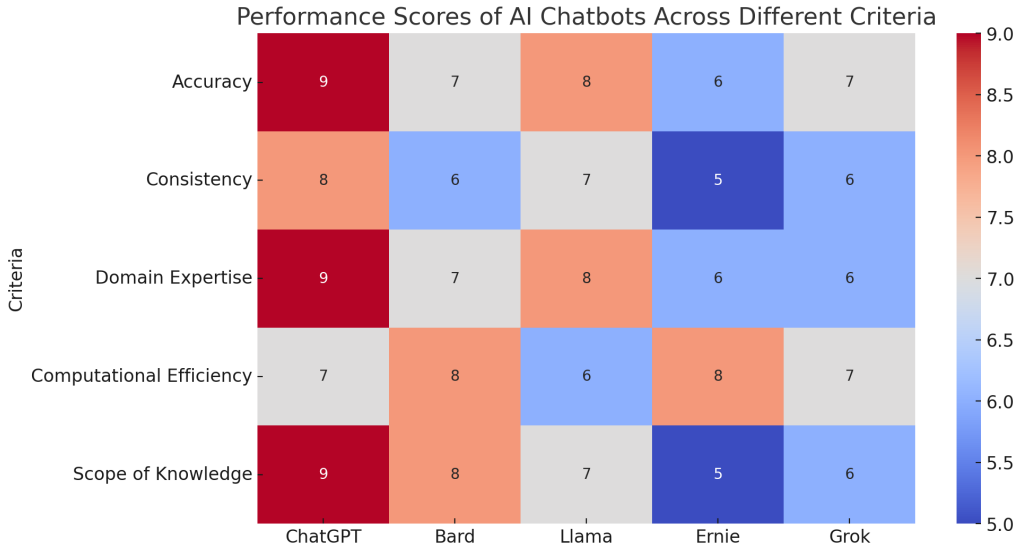

Figure 7. Heatmap: performance scores of AI chatbots across different criteria.

4.7.2. Insights from the Heatmap

The heatmap reveals several key insights into the performance of the AI chatbots:

- **ChatGPT**: Exhibited strong performance in accuracy, domain expertise, and scope of knowledge, thus making it a reliable tool for academic and practical tasks. However, its computational efficiency was only moderate.

- **Bard**: Excelled in real-time internet access and computational efficiency but had a lower consistency and scope of knowledge compared to ChatGPT.
- **Llama**: Showed balanced performance with scores ranging from 6 to 8 across all criteria, but it lacked standout strengths.
- **Ernie**: High computational efficiency was a notable strength, but it struggled with consistency and the scope of knowledge, thus limiting its usability outside its primary domain.
- **Grok**: Maintained moderate performance across all criteria, thus indicating versatility but not exceptional capability in any specific area.

This heatmap provides a comprehensive visual summary, thus aiding in the quick comparison of the AI chatbots' performance and helping to identify the most suitable tool for specific applications. It is important to note that these scores were not based on the experiments we conducted, but rather on the data and insights extracted from the 28 papers reviewed in this study. Assumptions were made to quantify these data points, thus offering general insights rather than precise representations of performance.

4.8. Real-World Applications of AI Chatbots

The comparative analysis and synthesis of the scholarly literature on AI chatbots, specifically ChatGPT, Bard, Llama, Ernie, and Grok, highlight their diverse capabilities and potential for real-world applications across various domains. This section explores some of the practical use cases where these AI models are making a significant impact, thus demonstrating their value beyond academic research.

4.8.1. Education and Learning

AI chatbots have shown remarkable potential in transforming education and learning across various disciplines and levels. ChatGPT, in particular, has been at the forefront of this educational revolution, with numerous case studies highlighting its impact on student engagement, learning outcomes, and research efficiency.

One of the most promising areas for AI chatbot applications is in the field of education. ChatGPT has shown strong potential as an educational tool, especially for novice learners. Its ability to generate multiple choice questions and evaluate educational materials, such as those related to sleep apnea [26], underscores its utility in creating and curating learning content. Moreover, ChatGPT 3.5's proficiency in programming across languages like Java, Python, and C++, as well as its adaptability to cater to both non-programmers and experienced coders [33], further highlights its potential to revolutionise coding education and practice.

In the domain of higher education English teaching, ChatGPT has demonstrated its ability to enhance language learning by providing instant feedback, generating diverse conversational scenarios, and facilitating language practice [34]. This interactive and personalised approach has led to improved student language skills and increased engagement in the learning process.

Beyond language learning, ChatGPT has also proven valuable in streamlining the academic research processes. The use of ChatGPT for thematic analysis showcases its potential to assist researchers in coding qualitative data, identifying themes, and drafting thematic reports, although it is not without challenges [35]. This human–AI collaboration significantly reduces the time and effort required for analysis, thus allowing researchers to focus on higher-level tasks.

The impact of ChatGPT on higher education extends further to provide assistance in tutoring, content creation, and administrative support [36]. Many related studies reported increase student satisfaction, learning efficiency, and an improvement to the overall educational experience when ChatGPT was integrated into the learning environment.

In experiential learning environments, GenAI tools like ChatGPT were also used to facilitate authentic assessment by providing realistic scenarios and personalised feed-

back [37]. This approach enhances students' critical thinking and problem-solving skills, thus preparing them for real-world challenges.

As AI chatbots continue to evolve and integrate into educational systems, their impact on learning and teaching practices is set to grow exponentially. As researchers, we have the responsibility to create methods that allow for the full exploration of the AI potentials, which could revolutionise traditional learning and teaching. This mainly includes the provisions of the necessary ethical and responsible use guidelines and practices.

4.8.2. Healthcare and Medicine

Another domain where AI chatbots are making significant advancement is healthcare. Llama's potential to develop a ChatDoctor, predict diagnosis-related groups, and integrate with natural language processing models like BERT [38] showcases AI's promising applications in medical knowledge dissemination and patient care. More studies on the application of AI in the medical domain have started to emerge such as this study, which assessed the sensitivity, specificity, and accuracy of ChatGPT 3.5, ChatGPT 4, Bing AI, and Bard in predicting drug–drug interactions [1,39], which underscores its transformative potential in healthcare education and practice.

4.8.3. Language and Cultural Diversity

AI chatbots are also making an impact in the domain of language and cultural diversity. Ernie's proficiency in comprehending various Chinese dialects and its integration with Baidu [4] positions it as a powerful tool for serving the linguistic diversity of China's population and becoming an integral part of the country's digital ecosystem. This highlights the potential for AI models to bridge language barriers and promote cultural understanding.

4.8.4. Social Media and User Engagement

Grok's promising capabilities in real-time social media insights, its engagement with "spicy" questions, and its incorporation of humor [30] set it apart from other LLMs, thus suggesting its potential for unique applications in social media analysis and user interaction. As AI chatbots continue to evolve, their ability to engage users on a more personal and interactive level could transform the landscape of social media and customer service.

The real-world applications of AI chatbots across education, healthcare, language and culture, and social media underscore their transformative potential. As these models continue to advance and address their limitations, such as accuracy, consistency, and ethical considerations, their impact on various domains is set to grow exponentially. The review of the scholarly literature in this study provides a foundation for understanding the current state of AI chatbots and their practical use cases, thus paving the way for future research and development in this dynamic field.

4.9. Reported Limitations from Scholarly Analysis

Despite the remarkable capabilities of each chatbot, scholarly analysis has revealed inherent limitations that necessitate careful consideration in their application. A prevalent concern across all models is the potential cybersecurity risks that they introduce [40]. THe inadvertent input of sensitive or classified information into these systems could expose such data to human reviewers or, more alarmingly, to cybercriminals who may exploit AI tools to orchestrate attacks, including hacking and phishing [41]. Furthermore, AI often faces scrutiny for its potential to be invasive, biased, and manipulative [42]. For instance, ChatGPT has been criticised for exhibiting a perceived political bias, which can distort its outputs [43]. These findings underscore the crucial importance of adhering to ethical principles to ensure secure, safe, and responsible AI deployment.

Focusing on ChatGPT, despite its proficiency in delivering educational content and medical information, it is not infallible. Inconsistencies in information accuracy and limitations in displaying graphical content have been identified as notable challenges [26]. Instances where ChatGPT provided inconsistent responses under identical conditions have

raised concerns about its reliability [24]. Additionally, academic integrity, context comprehension, and safety have been highlighted as areas requiring attention [44]. Significantly, neither ChatGPT 3.5 nor 4.0 has demonstrated the capability to generate original scientific contributions [39], and there have been documented cases of these models failing to accurately predict drug interactions [1]. In the educational domain, evidence suggests that the overreliance on ChatGPT 3.5 may lead to suboptimal academic performance [33]. Moreover, the knowledge base of these models is constrained to a specific point in time, thus limiting their ability to provide up-to-date information [38]. It is also noteworthy that, currently, both ChatGPT 3.5 and 4.0 are not open source, thus potentially restricting their accessibility and opportunities for collaborative enhancement.

Bard, while offering real-time internet access, has been observed to generate inconsistent answers and struggle with email consistency [27,45]. Its performance in anatomy writing and generating original scientific contributions has also been questioned [46]. Similarly, Llama's suitability for advanced mathematical topics like calculus and statistics has been found to be limited, and it requires substantial computational resources [26].

Ernie, being primarily focused on mainland China, may face challenges in terms of international accessibility and the impact of strict Chinese policies on its performance [4]. Its biennial knowledge updates also suggest potential limitations in staying current with rapidly evolving information.

Grok, although showing promise in real-time information processing and humour, is still in its early stages of development. Its current exclusivity to X premium subscribers and the presence of biased or inaccurate information are notable drawbacks [30]

Recognising these limitations, users and developers must approach the utilisation of AI models with a balanced understanding of their capabilities and an awareness of the potential risks. Prudent use of these technologies is essential to harness their benefits while mitigating their drawbacks. As the field of AI continues to evolve, ongoing research and development efforts should focus on addressing these limitations and enhancing the robustness, reliability, and ethical alignment of these powerful tools.

These are our suggestions for rewriting the section on the real-world applications of AI chatbots to better integrate them with the rest of the content available and to improve their flow and cohesiveness.

5. Study Limitations

This investigation, based on an analysis of 28 scholarly articles, has inherent limitations. The selected papers, while significant, represent only a fraction of the rapidly expanding field of AI language models. Consequently, the insights and comparative analyses derived herein are limited to the content and scope of these specific studies, thus potentially omitting a broader spectrum of capabilities and deficiencies inherent to these technologies.

Moreover, the study lacks original experimental data or empirical validation of the models' performances, thus making the findings reliant on the methodologies and conclusions of third-party researchers. The rapid evolution characteristic of AI development may also reduce the relevance of the conclusions as newer model iterations emerge. This highlights the critical need for continuous, updated evaluations to keep pace with advancements in the field.

Furthermore, the heterogeneity in methodologies and focal points across the reviewed papers could introduce variability, thus complicating direct comparisons. The primary reliance on English language sources further limits the generalisability of the findings, as language models operate in a multilingual global context. Future studies would benefit from incorporating a broader linguistic perspective to attain a more comprehensive understanding of these models' capabilities across different languages and cultural contexts.

In addition to these limitations, there are technical challenges and application-specific considerations that warrant discussion. For example, ensuring data security and preventing unauthorised access to sensitive information is a critical concern, particularly in healthcare

applications like Llama's ChatDoctor [38]. Implementing robust encryption protocols and conducting regular security audits are essential to mitigate these risks.

Moreover, the performance and applicability of AI chatbots may vary across different languages and cultures. While models like Ernie excel in comprehending Chinese dialects [4], their effectiveness in other linguistic contexts remains uncertain. Developing chatbots that can navigate cultural nuances and adapt to regional preferences is an ongoing challenge that requires further research and development.

Despite these limitations, this study provides a valuable foundation for future research by synthesising the current state of knowledge and identifying key areas for further investigation. As the field of AI language models continues to evolve, it is crucial to build upon these findings and conduct more extensive, empirical studies to validate and expand the understanding of these technologies.

6. Ethical and Privacy Issues

The rapid integration of AI chatbots into various sectors, particularly healthcare and education, raises significant ethical and privacy concerns that must be addressed to ensure responsible development and usage. This section discusses the impact of these issues and offers suggestions for mitigating the risks associated with AI chatbot applications.

6.1. Data Privacy and Security in Healthcare

In healthcare, AI chatbots like Llama's ChatDoctor handle sensitive patient data, thus making them targets for data breaches and unauthorised access. The ethical handling of medical information is governed by regulations such as HIPAA in the United States, GDPR in Europe, and the Australian Privacy Act of 1988 in Australia, which emphasize the need for stringent data protection measures. However, chatbots can inadvertently expose patient data through security vulnerabilities, as highlighted by the concerns raised in the evaluation of ChatGPT 4.0's performance in medical domains [1].

Recommendations

The following recommendations are proposed to enhance data privacy and security for AI chatbots:

- Implement end-to-end encryption for all data exchanges involving AI chatbots to safeguard patient information, as suggested in [47]. Extend privacy policies to govern AI use such as anonymisation, securing data at rest, and rights of data owners to access their data and delete them.
- Conduct regular security audits and updates to the chatbot systems to address potential vulnerabilities and ensure compliance with healthcare privacy laws, as outlined in the study of the security and privacy risks associated with AI agents in [48].
- Employ secure data storage and access control mechanisms to prevent unauthorised third-party access to sensitive patient information, as highlighted in [49].

6.2. Bias and Fairness in Educational Chatbots

AI chatbots in education, such as ChatGPT, must navigate issues of bias, which can manifest in the resources they provide or the interactions they facilitate. Biases in AI systems often stem from the datasets used for training these models, which can skew chatbot responses based on incomplete or unrepresentative data. This concern is evident in the evaluation of four AI chatbots for calculus and statistics educational materials in [26], which recommended that educators must be aware of the potential bias issue and to take steps to mitigate it, such as by ensuring that AI chatbots are trained on a diverse and inclusive dataset.

Recommendations

The following recommendations are proposed to address bias and promote fairness in educational chatbots:

- Diversify AI training datasets to include a wide range of demographics, experiences, and educational backgrounds to reduce bias.
- Regularly update and review the chatbot's decision-making algorithms to ensure that they remain unbiased and equitable, and upate legal and regulatory frameworks to keep pace with AI advancements and ensure their alignment with ethical principles and societal values, as pointed out in [50].
- Develop chatbots that can adapt to different cultural contexts and learning styles, as cultural differences may influence the effectiveness and reception of educational chatbots across regions. As the unwise use of has AI-enhanced, educational technology for learning and teaching may increase the structural disparities and digital divide already inherent in today society, as concluded in the study reported in [51].

6.3. Consent and Transparency in User Interactions

Transparency in how chatbots gather, use, and store user data is crucial for building trust. Users should be informed about the extent of the chatbot's capabilities, including its use of personal data and the limitations of its advice, particularly in sensitive fields such as medical and healthcare applications. Concerns similar to these have been raised in the context of higher education in India, where the adoption of AI chatbots has been scrutinised [52].

Recommendations

The following recommendations are proposed to address consent and transparency issues in User Interactions in chatbots used in sensitive applications:

- Clearly communicate to users how their data will be used, what data will be stored, and for how long through easily understandable privacy policies, as emphasised in [53], which stresses on the importance of exploring encryption secure protocols and authentication mechanisms to ensure the confidentiality and integrity of patient data.
- Implement user consent protocols, thus ensuring users that are fully aware of and agree to how their information is handled and whether it is used for training the AI model.
- Provide users with options to control their data, such as the ability to delete their conversation history or opt-out of data collection and data training.

6.4. Accountability in Decision Making

As chatbots become more autonomous, determining accountability for decisions made by these systems, especially erroneous ones that could lead to harm, becomes challenging. This is particularly critical in fields like healthcare, where decisions have significant consequences, as evidenced by the concerns raised in the evaluation of ChatGPT's ability to predict drug interactions [1].

Recommendations

The following recommendations are proposed to address accountability in decision-making by AI chatbots:

- Establish clear guidelines on the accountability of chatbot actions, particularly in cases where they operate semi-autonomously. For instance, some red flags were raised in the analysis of ChatGPT's performance in medical licensing examinations, as reported in [54].
- Ensure that there is always a human in the loop in critical decision-making processes, especially in sensitive, mission-critical applications and legal scenarios, to supervise and override chatbot decisions when necessary.
- Implement robust monitoring and auditing systems to track chatbot decisions and identify potential errors or biases, such as the efforts reported in [55].

Addressing these ethical and privacy issues is crucial for the responsible development and deployment of AI chatbots. By implementing robust data protection measures,

ensuring unbiased and fair interactions, maintaining transparency in user consent, and establishing clear accountability guidelines, we can harness the potential of these technologies while mitigating the risks associated with their use. Furthermore, considering the technical limitations and application challenges, such as data security, cultural adaptability, and language-specific performance, is essential for the effective and ethical integration of AI chatbots across various domains.

7. Conclusions

The comparative analysis of AI chatbots reveals differentiated strengths and weaknesses across various domains, thus highlighting the absence of a singularly superior model. This analysis spans sectors, including medicine, technology, education, ethics, humanities, and writing, thus showcasing the context-dependent utility of each model.

ChatGPT, built upon the Generative Pre-Trained Transformer architecture, excels in medical, educational, humanities, and writing applications. It demonstrates strong performance in accuracy, domain expertise, and the scope of knowledge. However, it exhibits limitations in providing real-time and consistent information, lacks open source flexibility, and demonstrates potential ideological bias.

Bard, recently replaced as Gemini, leverages its integration with Google Search to proficiently address queries derived from the internet, thereby excelling in real-time information access and computational efficiency. Both ChatGPT and Bard show potential for success in competitive quiz environments but may influence academic rigour differently due to the variability in their responses.

Llama emerges as a formidable medical chatbot with significant contributions to Natural Language Processing and personalised educational tools, although its extensive computational demands may restrict broader application. Ernie, tailored for Mainland China, offers localised benefits but may face limitations in international contexts due to restrictive policies.

Grok, still in its early stages, shows promise in providing engaging, real-time interactions and humor, alongside capabilities in mathematical reasoning. Its full potential remains to be evaluated as it undergoes further development and empirical testing.

The limitations of this study necessitate a cautious interpretation of its conclusions. Notably, the scope of the research on models like Ernie and Llama was not exhaustive, and the emerging status of models such as Grok invites further investigation. The dynamic nature of AI advancements, exemplified by recent developments in Grok and Bard, underscores the continuous evolution of the field. Additionally, the lack of original experimental data and the reliance on English language sources limit the generalisability of the findings, thus emphasising the need for more comprehensive and multilingual research.

Furthermore, this study highlights the critical importance of addressing the ethical and privacy concerns associated with AI chatbots. The potential for data breaches, biased interactions, the lack of transparency in user consent, and unclear accountability in decision making are significant challenges that must be addressed to ensure the responsible development and deployment of these technologies. Implementing robust data protection measures, diversifying training datasets, adapting chatbots to different cultural contexts, and establishing clear guidelines for accountability are crucial steps in mitigating these risks.

Future research should expand to include a wider range of fields. It should explore practical applications such as predictive analytics in elections or real estate, thus enhancing the empirical foundation and applicability of AI language models. Additionally, researchers should address the limitations identified in this study, such as the lack of real-time information, inconsistencies in responses, and potential biases. Collaborative efforts between developers, researchers, and educators [56,57] are essential to improve these technologies' robustness, reliability, and ethical alignment.

As AI language models continue to evolve and become more integrated into various domains, it is crucial to establish clear guidelines and regulations to ensure their responsible development and deployment. This includes addressing concerns related to data privacy, ethics, security, and the potential misuse of these technologies. By fostering interdisciplinary

collaboration, conducting rigorous empirical studies, and prioritizing ethical considerations, the potential of these technologies can be harnessed while mitigating their risks and limitations, thus ultimately shaping a future where AI language models serve as valuable tools for the betterment of society.

In conclusion, this paper aims to contribute to an ongoing scholarly dialogue rather than serve as a definitive assessment of these technologies. Researchers and practitioners are invited to build upon these preliminary findings to enrich the collective understanding and application of AI language models in diverse global contexts. By addressing the technical limitations, application challenges, and ethical considerations highlighted in this study, we can work towards the responsible and effective integration of AI chatbots across various domains, thus unlocking their potential to revolutionise the way we learn, communicate, and make decisions.

Author Contributions: K.W.: Conceptualisation, Writing—Original Draft Preparation. S.K.: Conceptualisation, Writing—Original Draft Preparation. E.G.: Supervision, Project Administration, Review and Editing, Funding Acquisition. M.E. (Corresponding Author): Data Curation, Writing—Final Draft, Review and Editing, Intelligent Data Visualisation, Insights, Recommendations, and Advanced Analysis. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding authors. The authors declare that they have not used Artificial Intelligence (AI) tools in the writing of this article. However, ChatGPT 4 was employed for specific formatting tasks, such as converting tables into LaTeX and enhancing the readability of these tables. ChatGPT 4 was also utilised for copyediting and proofreading certain sections, including the abstract and the conclusion. Additionally, ChatGPT 4 assisted in improving the flow between sections and identifying the limitations of the work. Furthermore, Whimsical, in combination with ChatGPT 4 and Claude 3 AI, was used to create the mindmap and heatmap displayed in Figures 6 and 7. The authors provided the data input and were in full control throughout the process; they also carefully reviewed the accuracy of the data represented in these visual aids. The authors took precautions to ensure that ChatGPT 4 and Claude AI did not introduce any text that was not authored by the original writers. All suggestions provided by ChatGPT 4 and Claude AI were thoroughly reviewed and revised by the authors to maintain the accuracy and integrity of the final manuscript. This process ensured that the content remained correct and faithful to the authors' original contributions. We hope this detailed acknowledgement contributes towards establishing a new ethical practice where the exact use of AI in academic research is transparently disclosed, as the integration of AI in research settings is imminent. We aim to encourage the responsible use and critical evaluation of AI tools to enhance the integrity and reliability of scholarly work.

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

References

1. Al-Ashwal, F.Y.; Zawiah, M.; Gharaibeh, L.; Abu-Farha, R.; Bitar, A.N. Evaluating the Sensitivity, Specificity, and Accuracy of ChatGPT-3.5, ChatGPT-4, Bing AI, and Bard Against Conventional Drug-Drug Interactions Clinical Tools. *Drug Healthc. Patient Saf.* **2023**, *15*, 137–147. [CrossRef] [PubMed]
2. Eggmann, F.; Weiger, R.; Zitzmann, N.U.; Blatz, M.B. Implications of large language models such as ChatGPT for dental medicine. *J. Esthet. Restor. Dent.* **2023**, *35*, 1098–1102. [CrossRef] [PubMed]
3. Rahaman, M.S.; Ahsan, M.M.T.; Anjum, N.; Rahman, M.M.; Rahman, M.N. The AI Race is on! Google's Bard and Openai's Chatgpt Head to Head: An Opinion Article. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4351785 (accessed on 19 June 2024).
4. Rudolph, J.; Tan, S.; Tan, S. War of the chatbots: Bard, Bing Chat, ChatGPT, Ernie and beyond. The new AI gold rush and its impact on higher education. *J. Appl. Learn. Teach.* **2023**, *6*, 364–389. [CrossRef]
5. Moons, P.; Van Bulck, L. Using ChatGPT and Google Bard to improve the readability of written patient information: A proof of concept. *Eur. J. Cardiovasc. Nurs.* **2023**, *23*, 122–126. [CrossRef]
6. Qin, H.; Ji, G.P.; Khan, S.; Fan, D.P.; Khan, F.S.; Gool, L.V. How Good is Google Bard's Visual Understanding? An Empirical Study on Open Challenges. *Mach. Intell. Res.* **2023**, *20*, 605–613. [CrossRef]
7. Agarwal, M.; Sharma, P.; Goswami, A. Analysing the applicability of ChatGPT, Bard, and Bing to generate reasoning-based multiple-choice questions in medical physiology. *Cureus* **2023**, *15*, e40977. [CrossRef] [PubMed]

8. Ali, R.; Tang, O.Y.; Connolly, I.D.; Fridley, J.S.; Shin, J.H.; Zadnik Sullivan, P.L.; Cielo, D.; Oyelese, A.A.; Doberstein, C.E.; Telfeian, A.E.; et al. Performance of ChatGPT, GPT-4, and Google Bard on a Neurosurgery Oral Boards Preparation Question Bank. *Neurosurgery* **2023**, *93*, 1090. [CrossRef]
9. Giannakopoulos, K.; Kavadella, A.; Salim, A.A.; Stamatopoulos, V.; Kaklamanos, E.G. Evaluation of the Performance of Generative AI Large Language Models ChatGPT, Google Bard, and Microsoft Bing Chat in Supporting Evidence-Based Dentistry: Comparative Mixed Methods Study. *J. Med. Internet Res.* **2023**, *25*, e51580. [CrossRef] [PubMed]
10. Ram, B.; Verma, P. Artificial intelligence AI-based Chatbot study of ChatGPT, Google AI Bard and Baidu AI. *World J. Adv. Eng. Technol. Sci.* **2023**, *8*, 258–261. [CrossRef]
11. Seth, I.; Lim, B.; Xie, Y.; Cevik, J.; Rozen, W.M.; Ross, R.J.; Lee, M. Comparing the Efficacy of Large Language Models ChatGPT, BARD, and Bing AI in Providing Information on Rhinoplasty: An Observational Study. *Aesthetic Surg. J. Open Forum* **2023**, *5*, ojad084. [CrossRef]
12. Collins, J.A.; Fauser, B.C. Balancing the strengths of systematic and narrative reviews. *Hum. Reprod. Update* **2005**, *11*, 103–104. [CrossRef] [PubMed]
13. Levy, Y.; Ellis, T.J. A Systems Approach to Conduct an Effective Literature Review in Support of Information Systems Research. *Informing Sci. Int. J. Emerg. Transdiscipl.* **2006**, *9*, 181–212. [CrossRef] [PubMed]
14. Bland, M.; Burke, M.I.; Bertolaccini, K. Taking steps toward healthy & sustainable transport investment: A systematic review of economic evaluations in the academic literature on large-scale active transport infrastructure. *Int. J. Sustain. Transp.* **2024**, *18*, 201–220. [CrossRef]
15. Karim, S.; Uddin, S.; Imam, T.; Moni, M.A. A Systematic Review of Network Studies Based on Administrative Health Data. *Int. J. Environ. Res. Public Health* **2020**, *17*, 2568. [CrossRef] [PubMed]
16. Nadeem, A.; Marjanovic, O.; Abedin, B. Gender bias in AI-based decision-making systems: A systematic literature review. *Australas. J. Inf. Syst.* **2022**, *26*. [CrossRef]
17. Kitchenham, B.; Brereton, O.P.; Budgen, D.; Turner, M.; Bailey, J.; Linkman, S. Systematic literature reviews in software engineering–a systematic literature review. *Inf. Softw. Technol.* **2009**, *51*, 7–15. [CrossRef]
18. Huang, J.; Wang, H.; Sun, Y.; Shi, Y.; Huang, Z.; Zhuo, A.; Feng, S. ERNIE-GeoL: A Geography-and-Language Pre-trained Model and its Applications in Baidu Maps. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, 14–18 August 2022; ACM: New York, NY, USA, 2022; pp. 3029–3039. [CrossRef]
19. Idrisov, B.; Schlippe, T. Program Code Generation with Generative AIs. *Algorithms* **2024**, *17*, 62. [CrossRef]
20. Jovic, M.; Mnasri, S. Evaluating AI-Generated Emails: A Comparative Efficiency Analysis. *World J. Engl. Lang.* **2024**, *14*, 502. [CrossRef]
21. Kumari, A.; Kumari, A.; Singh, A.; Singh, S.K.; Juhi, A.; Dhanvijay, A.K.D.; Pinjar, M.J.; Mondal, H. Large Language Models in Hematology Case Solving: A Comparative Study of ChatGPT-3.5, Google Bard, and Microsoft Bing. *Cureus* **2023**, *15*, e43861. [CrossRef]
22. Li, D.J.; Kao, Y.C.; Tsai, S.J.; Bai, Y.M.; Yeh, T.C.; Chu, C.S.; Hsu, C.W.; Cheng, S.W.; Hsu, T.W.; Liang, C.S.; et al. Comparing the performance of ChatGPT GPT-4, Bard, and Llama-2 in the Taiwan Psychiatric Licensing Examination and in differential diagnosis with multi-center psychiatrists. *Psychiatry Clin. Neurosci.* **2024**, *78*, 347–352. [CrossRef]
23. Meo, S.A.; Al-Khlaiwi, T.; AbuKhalaf, A.A.; Meo, A.S.; Klonoff, D.C. The Scientific Knowledge of Bard and ChatGPT in Endocrinology, Diabetes, and Diabetes Technology: Multiple-Choice Questions Examination-Based Performance. *J. Diabetes Sci. Technol.* **2023**, 19322968231203987. [CrossRef] [PubMed]
24. O'Leary, D.E. An analysis of Watson vs. BARD vs. ChatGPT: The Jeopardy! Challenge. *AI Mag.* **2023**, *44*, 282–295. [CrossRef]
25. Browne, R. Italy Became the First Western Country to Ban ChatGPT. Here's What Other Countries Are Doing. 2023. Available online: https://cacmb4.acm.org/news/271919-italy-became-the-first-western-country-to-ban-chatgpt-heres-what-other-countries-are-doing/fulltext (accessed on 22 May 2024).
26. Calonge, D.S.; Smail, L.; Kamalov, F. Enough of the chit-chat: A comparative analysis of four AI chatbots for calculus and statistics. *J. Appl. Learn. Teach.* **2023**, *6*, 346–357. [CrossRef]
27. Plevris, V.; Papazafeiropoulos, G.; Jiménez Rios, A. Chatbots Put to the Test in Math and Logic Problems: A Comparison and Assessment of ChatGPT-3.5, ChatGPT-4, and Google Bard. *AI* **2023**, *4*, 949–969. [CrossRef]
28. Insuasti, J.; Roa, F.; Zapata-Jaramillo, C.M. Computers' Interpretations of Knowledge Representation Using Pre-Conceptual Schemas: An Approach Based on the BERT and Llama 2-Chat Models. *Big Data Cogn. Comput.* **2023**, *7*, 182. [CrossRef]
29. Yu, C. PaddlePaddle/ERNIE. 2024. Available online: https://github.com/hotpads/ERNIE-for-the-Rest-of-Us (accessed on 19 June 2024).
30. Jin, H.; Dang, S. Elon Musk Says He Will Launch Rival to Microsoft-Backed ChatGPT | Reuters. 2023. Available online: https://www.reuters.com/technology/musk-says-he-will-start-truthgpt-or-maximum-truth-seeking-ai-fox-news-2023-04-17/ (accessed on 22 May 2024).
31. Xai. Open Release of Grok-1. 2024. Available online: https://x.ai/blog/grok-os (accessed on 22 May 2024).
32. Whimsical. Whimsical—Collaborative Online Diagramming Tool. 2024. Available online: https://whimsical.com/ (accessed on 21 May 2024).
33. Khademi, A. Can ChatGPT and Bard generate aligned assessment items? A reliability analysis against human performance. *J. Appl. Learn. Teach.* **2023**, *6*, 75–80. [CrossRef]

34. Kostka, I.; Toncelli, R. Exploring applications of ChatGPT to English language teaching: Opportunities, challenges, and recommendations. *TESL-EJ* **2023**, *27*, n3. [CrossRef]
35. Lee, V.V.; van der Lubbe, S.C.; Goh, L.H.; Valderas, J.M. Harnessing ChatGPT for thematic analysis: Are we ready? *J. Med. Internet Res.* **2024**, *26*, e54974. [CrossRef] [PubMed]
36. Vargas-Murillo, A.R.; de la Asuncion, I.N.M.; de Jesús Guevara-Soto, F. Challenges and opportunities of AI-assisted learning: A systematic literature review on the impact of ChatGPT usage in higher education. *Int. J. Learn. Teach. Educ. Res.* **2023**, *22*, 122–135. [CrossRef]
37. Salinas-Navarro, D.E.; Vilalta-Perdomo, E.; Michel-Villarreal, R.; Montesinos, L. Designing experiential learning activities with generative artificial intelligence tools for authentic assessment. *Interact. Technol. Smart Educ.* 2024, *ahead-of-print*. [CrossRef]
38. Zandi, R.; Fahey, J.D.; Drakopoulos, M.; Bryan, J.M.; Dong, S.; Bryar, P.J.; Bidwell, A.E.; Bowen, R.C.; Lavine, J.A.; Mirza, R.G. Exploring Diagnostic Precision and Triage Proficiency: A Comparative Study of GPT-4 and Bard in Addressing Common Ophthalmic Complaints. *Bioengineering* **2024**, *11*, 120. [CrossRef]
39. Lozić, E.; Štular, B. Fluent but Not Factual: A Comparative Analysis of ChatGPT and Other AI Chatbots' Proficiency and Originality in Scientific Writing for Humanities. *Future Internet* **2023**, *15*, 336. [CrossRef]
40. Humphreys, D.; Koay, A.; Desmond, D.; Mealy, E. AI hype as a cyber security risk: The moral responsibility of implementing generative AI in business. *AI Ethics* **2024**. [CrossRef]
41. Gupta, M.; Akiri, C.; Aryal, K.; Parker, E.; Praharaj, L. From ChatGPT to ThreatGPT: Impact of Generative AI in Cybersecurity and Privacy. *IEEE Access* **2023**, *11*, 80218–80245. [CrossRef]
42. Schlagwein, D.; Willcocks, L. 'ChatGPT et al.': The ethics of using (generative) artificial intelligence in research and science. *J. Inf. Technol.* **2023**, *38*, 232–238. [CrossRef]
43. Teubner, T.; Flath, C.M.; Weinhardt, C.; van der Aalst, W.; Hinz, O. Welcome to the Era of ChatGPT et al. *Bus. Inf. Syst. Eng.* **2023**, *65*, 95–101. [CrossRef]
44. Cotton, D.R.E.; Cotton, P.A.; Shipway, J.R. Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innov. Educ. Teach. Int.* **2023**, *61*, 228–239. [CrossRef]
45. Patil, N.S.; Huang, R.S.; van der Pol, C.B.; Larocque, N. Comparative Performance of ChatGPT and Bard in a Text-Based Radiology Knowledge Assessment. *Can. Assoc. Radiol. J.* **2023**, *75*, 344–350. [CrossRef]
46. Ilgaz, H.B.; Çelik, Z. The significance of artificial intelligence platforms in anatomy education: An experience with ChatGPT and google bard. *Cureus* **2023**, *15*, e45301. [CrossRef]
47. Yang, J.; Chen, Y.L.; Por, L.Y.; Ku, C.S. A systematic literature review of information security in chatbots. *Appl. Sci.* **2023**, *13*, 6355. [CrossRef]
48. Falade, P.V. Investigating The Security and Privacy Issues in ChatGPT Usage and Their Impact on Organisational and Individual Security. *Int. J. Sci. Res. Multidiscip. Stud.* **2024**, *10*, 19–30.
49. Gabriel, O.T. Data Privacy and Ethical Issues in Collecting Health Care Data Using Artificial Intelligence Among Health Workers. Master's Thesis, Center for Bioethics and Research, Ibadan, Nigeria, 2023.
50. Williamson, S.M.; Prybutok, V. The Era of Artificial Intelligence Deception: Unraveling the Complexities of False Realities and Emerging Threats of Misinformation. *Information* **2024**, *15*, 299. [CrossRef]
51. Dieterle, E.; Dede, C.; Walker, M. The cyclical ethical effects of using artificial intelligence in education. *AI Soc.* **2024**, *39*, 633–643. [CrossRef] [PubMed]
52. Sandu, N.; Gide, E. Adoption of AI-Chatbots to Enhance Student Learning Experience in Higher Education in India. In Proceedings of the 2019 18th International Conference on Information Technology Based Higher Education and Training (ITHET), Magdeburg, Germany, 26–27 September 2019; pp. 1–5. [CrossRef]
53. Garcia Valencia, O.A.; Suppadungsuk, S.; Thongprayoon, C.; Miao, J.; Tangpanithandee, S.; Craici, I.M.; Cheungpasitporn, W. Ethical implications of chatbot utilization in nephrology. *J. Pers. Med.* **2023**, *13*, 1363. [CrossRef] [PubMed]
54. Le, M.; Davis, M. ChatGPT Yields a Passing Score on a Pediatric Board Preparatory Exam but Raises Red Flags. *Glob. Pediatr. Health* **2024**, *11*, 2333794X241240327. [CrossRef] [PubMed]
55. Koshiyama, A.; Kazim, E.; Treleaven, P.; Rai, P.; Szpruch, L.; Pavey, G.; Ahamat, G.; Leutner, F.; Goebel, R.; Knight, A.; et al. Towards algorithm auditing: Managing legal, ethical and technological risks of AI, ML and associated algorithms. *R. Soc. Open Sci.* **2024**, *11*, 230859. [CrossRef]
56. Elkhodr, M.; Gide, E.; Wu, R.; Darwish, O. ICT students' perceptions towards ChatGPT: An experimental reflective lab analysis. *STEM Educ.* **2023**, *3*, 70–88. [CrossRef]
57. Sandu, R.; Gide, E.; Elkhodr, M. The role and impact of ChatGPT in educational practices: Insights from an Australian higher education case study. *Discov. Educ.* **2024**, *3*, 71. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.