

Assessing the Feasibility of Large Language Models to Identify Top Research Priorities in Enhanced External Counterpulsation

Shengkun Gai^{1,2}, Fangwan Huang³, Xuanyun Liu³, Ryan G. Benton⁴, Glen M. Borchert⁵, Jingshan Huang^{6□}, Xiuyu Leng^{1*}

¹ Department of Cardiology, The First Affiliated Hospital, Sun Yat-sen University, Guangzhou, China

² Department of Cardiology, Linfen People's Hospital, Shanxi, China

³ College of Computer and Data Science, Fuzhou University, Fuzhou, China

⁴ School of Computing, University of South Alabama, Mobile, AL, U.S.A.

⁵ College of Medicine, University of South Alabama, Mobile, AL, U.S.A.

⁶ School of Computing and College of Medicine, University of South Alabama, Mobile, AL, U.S.A.

□ Co-Corresponding author: huang@southalabama.edu

* Corresponding author: lengxy@mail.sysu.edu.cn

Abstract. Enhanced External Counterpulsation (EECP), as a non-invasive, cost-effective, and efficient adjunctive circulatory technique, has found increasingly widespread applications in the cardiovascular field. Numerous basic research and clinical observations have extensively confirmed the significant advantages of EECP in promoting blood flow perfusion to vital organs such as the heart, brain, and kidneys. However, many potential mechanisms of EECP remain insufficiently validated, necessitating researchers to dedicate substantial time and effort to in-depth investigations. In this work, we attempted to use large language models (such as ChatGPT and Ernie Bot) to identify top research priorities in five key topics in the field of EECP: mechanisms, device improvements, cardiovascular applications, neurological applications, and other applications. After generating specific research priorities in each domain through language models, we invited a panel of nine experienced EECP experts to independently evaluate and score them based on four parameters: relevance, originality, clarity, and specificity. Notably, average and median scores for these evaluation parameters were high, indicating a strong endorsement from experts in the EECP field. Although further validation and refinement are required, this study preliminarily suggests that large language models like ChatGPT and Ernie Bot could serve as powerful tools for identifying and prioritizing research priorities in the EECP domain.

Keywords: EECP, Research Priorities, Large Language Models, ChatGPT, Ernie Bot, Artificial Intelligence

1 INTRODUCTION

Enhanced External Counterpulsation (EECP) is a non-invasive adjunctive circulatory technique that inflates and deflates cuffs wrapped around the limbs and buttocks in sync with the cardiac cycle under electrocardiographic gating control. EECP has been shown clinically to significantly improve organ perfusion, regulate endothelial function, combat coronary artery atherosclerosis, treat complications of diabetes and

sudden sensorineural hearing loss, among other benefits, through a series of mechanisms [1-3]. While evidence suggests there is still a great deal of untapped potential for external counterpulsation, traditional approaches to identifying research priorities for EECP rely mainly on expert opinion and consensus building which are often labor-intensive and biased. In recent years, natural language processing (NLP) technology [4] has been increasingly recognized as a new means of identifying research priorities. Large language models (LLMs) such as ChatGPT [6] and Ernie Bot [7], which are trained on extensive text data, possess the ability to understand human-like language and have demonstrated significant potential in proposing and prioritizing research priorities [5]. In this work, ChatGPT and Ernie Bot were evaluated for their effectiveness in identifying primary research priorities related to EECP technology. Five key areas were examined: mechanisms, device enhancements, cardiovascular applications, neurological applications, and other applications. Utilizing ChatGPT and Ernie Bot, specific research priorities in these domains were generated after which experienced EECP experts reviewed and then rated them to assess their relevance and importance.

2 RELATED WORK

Large language models have shown broad applicability in entertainment, education, and customer service, but their potential in the medical field remains largely untapped. Given the high standards for information quality and communication reliability in medicine, the application of large language models requires careful consideration. In recent years, scholars have begun to explore the use of large language models in medicine, yielding promising results. In the field of cardiology, Gala et al. [8] believe that LLMs can analyze a large number of research papers and medical record resources to help clinicians keep up with the latest advances in cardiology. But they also point to the limitations of LLMs in explaining cultural or emotional factors that may influence medical practice. Cascella et al. [9] explored ChatGPT's reasoning abilities on public health topics. Through a question-and-answer session, ChatGPT listed four possible research topics. While some of ChatGPT's responses may be stereotyped and depend on the prompts, it can be used to summarize the scientific literature and generate new research hypotheses. Additionally, George et al. [10] proposed that large language models can serve as a supplementary resource to traditional medical tools, improving the efficiency and productivity of medical practices. Unfortunately, these studies do not provide a quantitative assessment of the LLMs' ability to identify medical research priorities.

Importantly, in order to assess the effectiveness of LLMs in the medical domain, it is essential to conduct statistical analyses on numerical results obtained from experiments and/or surveys. In evaluating the pertinent literature on LLMs, Tang et al. [11] invited field experts to assess the summary quality of LLMs by using a five-point Likert scale along four dimensions: coherence, factual consistency, comprehensiveness, and harmfulness. The Man-Whitney U test was used to assess the differences in response between GPT-3.5 and ChatGPT. Michael et al. [12] employed average scoring and fixed-effects consistency to calculate the Intraclass Correlation Coefficient (ICC), investigating the potential application of artificial intelligence-based LLMs in the realm of medical ethics. Similarly, Dave et al. [13] utilized Pearson and Spearman coefficients to juxtapose the assessment outcomes of large language models against the evaluations

of medical professionals, thereby further substantiating their dependability. Furthermore, besides correlation analysis, similarity metrics are frequently utilized to gauge the efficacy of LLMs. For example, in 2024, Sebastian et al. [14] evaluated the pairwise accuracy between LLMs and human assessments by analyzing the cosine similarity matrix. In measuring factual knowledge within LLMs, Pezeshkpour [15] successfully utilized Kullback-Leibler (KL) divergence to analyze the predictive probability distributions of the model before and after instilling target knowledge. Guo et al. [16], in investigating bias issues within large pre-trained language models, used the Jensen-Shannon (JS) divergence to measure the consistency between different demographic distributions, offering a robust tool for reducing human-like biases and unwanted societal stereotypes. JS divergence is an improved version of KL divergence, being symmetric whereas KL divergence is asymmetric, rendering JS divergence more precise in discerning similarity.

3 METHODS

Research priorities. We leveraged ChatGPT (based on GPT-3.5) which has captured 100 million users worldwide and Ernie Bot 3.5 which more Popular in China to generate research priorities in five key topics (Table 1 and Table 2, respectively) pertaining to EECF mechanisms [1], [17], structural enhancements, applications in cardiovascular domains [3], [18], [19], neurological applications [20], [21], and other applications [3], [22], [23].

Table 1. ChatGPT-generated research priorities on five key topics in the field of EECF research

Topic	ChatGPT-generated Research Priorities
Mechanisms	<ol style="list-style-type: none"> 1. Integrated Insights into EECF Effects: Unveiling Hemodynamic, Neuroendocrine, and Inflammatory Responses and Their Implications for Cardiovascular and Systemic Health. 2. How does EECF compare to other treatment modalities such as pharmacotherapy, invasive procedures, or lifestyle interventions? 3. Optimizing EECF Treatment: Investigating Timing and Frequency for Varied Patient Populations and Clinical Conditions. 4. Refining Patient Selection for EECF: Optimizing Outcomes Based on Demographics, Comorbidities, and Treatment Responsiveness. 5. Assessing the Long-Term Impact of EECF Therapy: Prospective Studies on Cardiovascular Outcomes, Quality of Life, and Healthcare Utilization.
Device improvements	<ol style="list-style-type: none"> 1. Revolutionizing EECF Treatment: Evaluating Compact and Portable Devices for Home-Based and Ambulatory Care, Integrating Technological Advancements for Enhanced Efficacy. 2. Advancements in EECF Technology: Improving Treatment Efficacy and Patient Experience through Compact and Portable Devices, Automated Pressure Control, and Real-Time Monitoring. 3. Enhancing Patient Experience in EECF Treatment: Investigating Ergonomic Enhancements for Adjustable Seating, Cushioning, and Noise Reduction. 4. Integrating EECF Devices with Telemedicine and Mobile Health: Enhancing Remote Monitoring and Patient Care. 5. Optimizing EECF Treatment: Investigating Safety Mechanisms for Enhanced Patient Security
In the field of heart disease	<ol style="list-style-type: none"> 1. Unlocking Vascular Relaxation: Investigating Signaling Pathways in EECF Treatment through eNOS, cGMP, and Vasodilatory Factors.

	<p>2. Unveiling the Future of Cardiac Health: Investigating EECF Treatment Integration with Telemedicine and Mobile Health for Remote Monitoring.</p> <p>3. Longitudinal Assessment of EECF Impact on Endothelial Function, Progenitor Cell Activity, and Vascular Inflammation: Implications for Cardiovascular Disease Prevention and Management.</p> <p>4. Unraveling Dynamic Hemodynamic Responses: Investigating Cardiac Output, Systemic Vascular Resistance Changes, and Implications for Non-Pharmacological Heart Failure Management in EECF Treatment.</p> <p>5. Examining Autonomic Regulation: HRV Parameters, Sympathetic-Parasympathetic Balance, and Baroreflex Sensitivity Dynamics Following EECF Treatment</p>
Applications in the field of neurology	<p>1. Neuroimaging Insights into Structural and Functional Changes Induced by EECF: Implications for Neural Repair and Recovery Processes.</p> <p>2. Synergistic Approaches to Neurological Rehabilitation: Investigating the Combined Effects of EECF with Physical, Occupational, and Cognitive Interventions on Motor and Cognitive Recovery Post-Stroke.</p> <p>3. Enhancing Neurological Recovery: Assessing the Impact of EECF on Cognitive Function, Neuroplasticity, and Rehabilitation Potential Following Traumatic Brain Injury.</p> <p>4. Exploring the Neurotrophic and Neuroregenerative Potential of EECF: Implications for Disease Pathology, Cognitive Decline, and Motor Symptoms.</p> <p>5. The Impact of Enhanced External Counterpulsation on Neurovascular Coupling and Cerebral Perfusion: Investigating Mechanisms and Potential Therapeutic Applications in Neurovascular Disorders.</p>
Applications in other fields	<p>1. Unveiling the Metabolic Impact of EECF: Exploring Effects on Insulin Sensitivity, Lipid Profiles, and Inflammatory Markers for Managing Metabolic Syndrome and Cardiovascular Risk Factors.</p> <p>2. Optimizing Orthopedic Wellness: Evaluating the Impact of EECF on Joint Function, Muscle Strength, and Pain Management in Osteoarthritis, Rheumatoid Arthritis, and Musculoskeletal Injuries.</p> <p>3. Enhancing Sleep Quality with EECF: Exploring Effects on Sleep Architecture, Latency, and Efficiency as a Non-Pharmacological Intervention for Sleep Disorders</p> <p>4. Renal Rejuvenation: Investigating the Effects of EECF on Renal Blood Flow, Glomerular Filtration Rate, and Kidney Injury Markers for Managing Renal Artery Stenosis and Acute Kidney Injury.</p> <p>5. Utilizing Enhanced External Counterpulsation to Enhance Cancer Treatment Outcomes: Investigating Chemotherapy Delivery, Cardiotoxicity Reduction, and Treatment Efficacy Enhancement.</p>

Table 2. ERNIE bot-generated research priorities on five key topics in the field of EECF research

Topic	ERNIE Bot-generated Research Priorities
Mechanisms	<p>1. Study on the impact of enhanced external counterpulsation on hemodynamics and collateral vessel development.</p> <p>2. Molecular mechanisms regulating endothelial cell function and smooth muscle cell behavior in EECF.</p> <p>3. Exploration of the mechanism of enhanced external counterpulsation based on genomics.</p> <p>4. Application of real-time imaging techniques in assessing the therapeutic effects of enhanced external counterpulsation.</p> <p>5. Comparative study of the efficacy of enhanced external counterpulsation and traditional drug therapy in cardiovascular and cerebrovascular diseases</p>
Device improvements	<p>1. Application of intelligent perception and adaptive adjustment technology in enhanced external counterpulsation machines.</p> <p>2. Optimization design of enhanced external counterpulsation machines based on biomechanical simulation.</p> <p>3. Research and application of new materials and technologies in enhanced external counterpulsation machines.</p> <p>4. Development and preliminary evaluation of wearable enhanced external counterpulsation devices.</p>

	5. Research on multimodal feedback and control systems in enhanced external counterpulsation machines.
In the field of heart disease	<ol style="list-style-type: none"> 1. Effect prediction and optimization of enhanced external counterpulsation in cardiovascular treatment based on big data analysis. 2. Exploration of the potential of enhanced external counterpulsation in post-heart transplant rehabilitation. 3. Mechanism study using enhanced external counterpulsation to promote heart regeneration and repair. 4. Exploration of personalized treatment strategies using enhanced external counterpulsation in cardiovascular medicine. 5. Long-term monitoring and management research in cardiovascular medicine using wearable enhanced external counterpulsation technology.
Applications in the field of neurology	<ol style="list-style-type: none"> 1. Mechanism of enhanced external counterpulsation in cerebral blood flow reconstruction in ischemic stroke patients. 2. Evaluation of the cognitive function improvement effect of enhanced external counterpulsation based on neuroimaging. 3. Promotion of brain function reshaping in neurorehabilitation through enhanced external counterpulsation. 4. Prospective study using enhanced external counterpulsation to improve the quality of life in neuropathic pain patients. 5. Synchronization study of brain blood flow and neural activity using enhanced external counterpulsation technology.
Applications in other fields	<ol style="list-style-type: none"> 1. Exploration of the mechanism of skin regeneration and wound healing using enhanced external counterpulsation technology. 2. Application of enhanced external counterpulsation in improving renal function in chronic kidney disease patients. 3. Promotion of healing and vascular regeneration in diabetic foot ulcers using enhanced external counterpulsation. 4. Potential study of enhanced external counterpulsation in improving joint blood flow and inflammation control in rheumatoid arthritis patients. 5. Role of enhanced external counterpulsation in promoting skin regeneration and reducing scar formation after burns.

Expert evaluation. The expert evaluation panel was comprised of nine highly experienced EECF specialists as evidenced by panelists having authored an average of twenty relevant research publications in the field. Panelists reviewed and assessed the research inquiries presented by ChatGPT and Ernie Bot independently. Experts rated five priorities on four parameters (relevance, originality, clarity, and specificity) using a 1-5 scale with 5 representing the highest score. ChatGPT and Ernie Bot-generated priorities were then compared to current EECF research queries identified through manual literature review. Importantly, in order to ensure the objectivity and relevance of responses, ChatGPT and Ernie Bot were instructed to treat each key topic as an independent query, thereby eliminating potential biases that may have existed in previous conversations.

4 STATISTICAL ANALYSIS

Standard statistical methods were utilized for both data collection and analysis with all statistical analyses carried out using IBM SPSS Statistics version 25 and Python 3.10. Initially, descriptive statistical methods were employed to provide a summary of the data, including measures such as mean, standard deviation (SD), and median. Following this, the study utilized "divergence" to assess the similarity between ratings provided by experts in EECF and queries generated by two large language models. In the realm of data mining, JS divergence was computed to evaluate the similarity of ratings among evaluators using a rating table structured with evaluators as column attributes. JS divergence values from 0 to 1, with smaller values indicating greater similarity between ratings. Additionally, Spearman's rank correlation coefficient and Kendall's τ

coefficient were also used to evaluate pairwise correlations between parameters. Positive coefficients indicate a positive correlation, while negative coefficients imply a negative correlation. The closer the coefficient is to 1 the stronger the correlation.

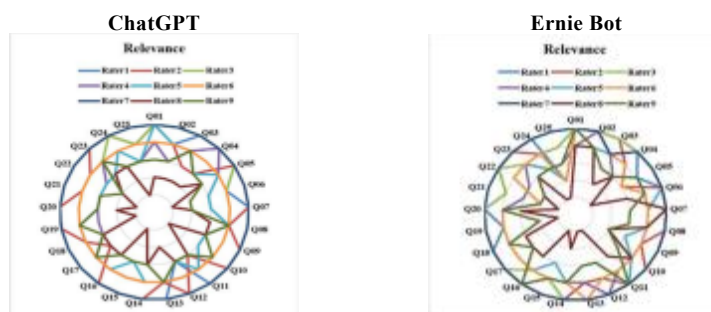
5 RESULTS

The statistical analysis shows high reliability for the questionnaires assessing ChatGPT and Ernie Bot, with Cronbach's alpha coefficients of 0.978 and 0.971, respectively. Both coefficients exceed the 0.8 threshold, indicating strong survey reliability. This suggests that the questionnaires effectively reflect the proficiency of ChatGPT and Ernie Bot in determining research priorities for EECF.

Based on this, the study conducted data analysis on the ratings provided by the 9 evaluators from three perspectives: (1) descriptive statistics; (2) similarity of ratings among evaluators; and (3) rank correlation of evaluation metrics. The data analysis tools utilized were IBM SPSS Statistics Version 25 and Python 3.10.

5.1 Descriptive statistics

In-depth descriptive statistical analyses of evaluation metrics are presented in Tables 3 to 5. The major models performed best in relevance, with originality close behind. While originality exhibited the largest standard deviation, suggesting significant variation in expert opinions regarding originality, clarity demonstrated the smallest standard deviation, indicating minimal fluctuations in scores for each question. Additionally, variations in performance between the two models (ChatGPT and Ernie Bot) across different evaluation metrics and topics can be observed. Concerning relevance, Ernie Bot's average score slightly exceeds ChatGPT's, suggesting a slight advantage in addressing user-related questions, although this was not statistically significant. In terms of originality, ChatGPT's score was slightly less than Ernie Bot's, with a higher fluctuation in scoring standard deviation, indicating some disagreement among experts regarding the originality of ChatGPT's queries. Both models demonstrate similar performance in clarity and specificity, indicating their similarity in providing clear and specific answers. Results of scores from EECF experts for all priorities are visually presented in Figure 1 with the outermost rings corresponding to the highest score of 5 and inner rings indicating lower scores.



Evaluating the Use of LLM to Identify Top Research Priorities in EECF

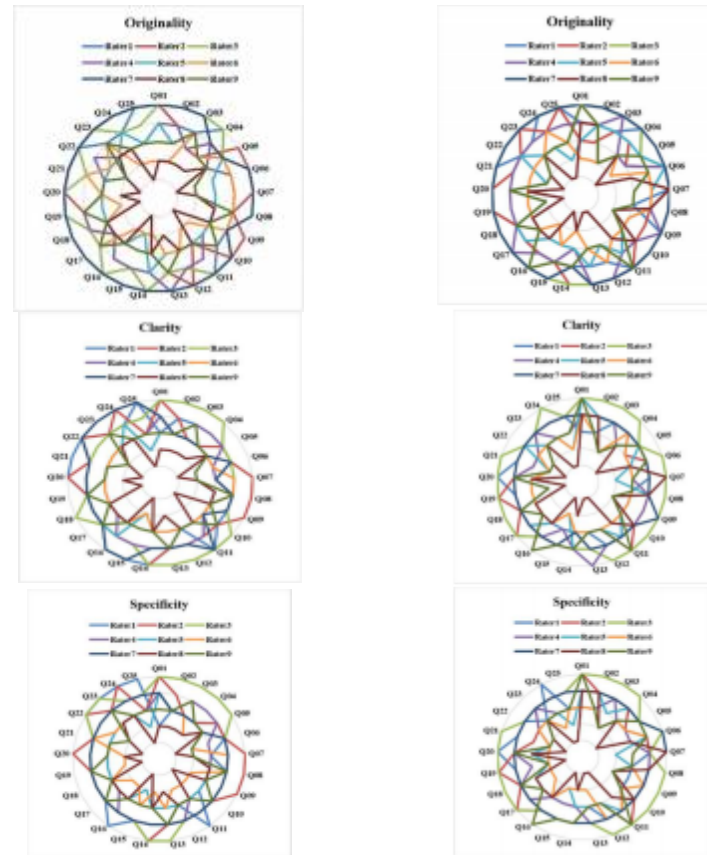


Figure 1. Ratings of 25 research focal points by nine evaluators based on four criteria.

Table 4 presents the scores given by different raters for the ChatGPT and Ernie Bot models. The analysis shows that in the evaluations of most raters, ChatGPT and Ernie Bot have similar average scores indicating a certain level of competitiveness in overall performance. However, it is worth noting that in the ratings of Rater3 and Rater4, Ernie Bot's average score was significantly higher than ChatGPT's, reflecting a more outstanding performance of Ernie Bot from the perspectives of these two raters. In terms of score stability, there were differences between the two models among different raters. Specifically, in the evaluations of Rater3 and Rater4, Ernie Bot had a lower standard deviation, indicating more stable scores and consistent performance. Conversely, Rater8's Ernie Bot scores demonstrated significantly higher standard deviation. In contrast, ChatGPT's standard deviation among multiple raters was relatively more consistent, although overall score stability was slightly inferior to Ernie Bot's performance for a subset of raters. These differences in evaluation may stem from personal preferences, evaluation criteria, and model performance across different topics.

In all topics (refer to Table 5), Ernie Bot consistently received higher average scores than ChatGPT, suggesting a relative advantage in overall performance. Although their performances in terms of median scores were similar, Ernie Bot achieved an upper quartile score of 5.00 in specific topics such as mechanisms, device improvements and applications in neurology, indicating higher recognition in these areas. Meanwhile ChatGPT's standard deviation across multiple topics was slightly lower than Ernie Bot's, suggesting relatively better score stability. However, this difference was not significant. Notably, clear domain-specific differences were observed, while Ernie Bot's average score significantly surpassed ChatGPT's in structural improvements and applications in neurology domains, ChatGPT demonstrated superior performance in other domains.

Table 3. Descriptive statistics of evaluation metrics

Evaluation Metric	Model	Mean	Standard Deviation	Lower Quartile	Median	Upper Quartile
relevance	ChatGPT	3.88	0.99	3.00	4.00	5.00
	Ernie Bot	4.04	1.06	4.00	4.00	5.00
originality	ChatGPT	3.70	1.05	3.00	4.00	5.00
	Ernie Bot	3.88	1.10	3.00	4.00	5.00
clarity	ChatGPT	3.52	0.92	3.00	4.00	4.00
	Ernie Bot	3.56	0.95	3.00	4.00	4.00
specificity	ChatGPT	3.40	0.93	3.00	3.00	4.00
	Ernie Bot	3.46	0.97	3.00	4.00	4.00

Table 4. Descriptive statistics of evaluator

Evaluator	Model	Mean	Standard Deviation	Lower Quartile	Median	Upper Quartile
Rater1	ChatGPT	4.34	0.71	4.00	4.00	5.00
	Ernie Bot	4.10	0.78	4.00	4.00	5.00
Rater2	ChatGPT	4.30	0.70	4.00	4.00	5.00
	Ernie Bot	4.20	0.67	4.00	4.00	5.00
Rater3	ChatGPT	4.37	0.49	4.00	4.00	5.00
	Ernie Bot	4.75	0.44	4.25	5.00	5.00
Rater4	ChatGPT	3.50	0.54	3.00	3.00	4.00
	Ernie Bot	3.81	0.61	3.00	4.00	4.00
Rater5	ChatGPT	3.16	0.58	3.00	3.00	3.00
	Ernie Bot	3.37	0.86	3.00	3.00	4.00
Rater6	ChatGPT	3.17	0.77	3.00	3.00	4.00
	Ernie Bot	3.14	0.79	3.00	3.00	4.00
Rater7	ChatGPT	4.44	0.67	4.00	5.00	5.00
	Ernie Bot	4.50	0.52	4.00	5.00	5.00
Rater8	ChatGPT	2.20	0.80	2.00	2.00	3.00
	Ernie Bot	2.44	1.21	1.00	3.00	3.00
Rater9	ChatGPT	3.16	0.61	3.00	3.00	4.00
	Ernie Bot	3.31	0.92	3.00	3.00	4.00

Table 5. Descriptive statistics of topic

topic	Model	Mean	Standard Deviation	Lower Quartile	Median	Upper Quartile
mechanisms	ChatGPT	3.63	0.95	3.00	4.00	4.00
	Ernie Bot	3.83	0.98	3.00	4.00	5.00
device improvements	ChatGPT	3.52	1.03	3.00	4.00	4.00
	Ernie Bot	3.94	0.90	3.00	4.00	5.00
in the field of heart disease	ChatGPT	3.69	0.98	3.00	4.00	4.00
	Ernie Bot	3.68	1.12	3.00	4.00	4.75
in the field of neurology	ChatGPT	3.67	0.91	3.00	4.00	4.00
	Ernie Bot	3.85	0.99	3.00	4.00	5.00
the otherfield	ChatGPT	3.62	1.06	3.00	4.00	4.00
	Ernie Bot	3.38	1.14	3.00	3.00	4.00

5.2 Similarity of raters' scores

Regarding the similarity of raters' scores, we calculated the JS divergence of scores between each pair of raters for ChatGPT and Ernie Bot (see Figure 2 for details). The results indicate that the JS divergence range of scores for ChatGPT is [0, 0.102], while for Ernie Bot, it is [0, 0.148]. Since a smaller JS divergence value indicates higher similarity, it can be concluded that the evaluations of these two large language models by raters exhibit relatively high consistency. It is worth noting that, for both ChatGPT and Ernie Bot, the similarity of scores between rater 8 and other raters is the lowest. From Figure 1, it is evident that the scores given by rater 8 are significantly lower than those given by other raters. Further analysis of the data in Table 4 reveals that the average scores given by rater 8 for both ChatGPT and Ernie Bot are the lowest (2.20 and 2.44 respectively), and they have the highest standard deviations (0.80 and 1.21 respectively). Excluding the influence of rater 8's scores, the upper limit of the JS divergence of scores for ChatGPT would decrease from 0.102 to 0.052, and from 0.148 to 0.063 for Ernie Bot.

5.3 Correlation of evaluation metrics

In terms of the correlation of evaluation metrics, we calculated both the Spearman [24] and Kendall [25] coefficients between pairs of evaluation metrics in the scoring results for ChatGPT and Ernie Bot (see Tables 6 and 7). These analyses passed significance tests, with all p-values below 0.01 indicating a significant positive correlation between relevance, originality, clarity, and specificity. This implies that when evaluating these two models, the score trends among these metrics were consistent, demonstrating high consistency and reliability. That said, ChatGPT exhibited a lower correlation between originality and relevance, while Ernie Bot showed a lower correlation in the analysis of specificity and relevance. The clarity of both models was highly correlated with relevance and/or specificity.

Table 6. Rank correlation coefficients between evaluation metrics (ChatGPT)

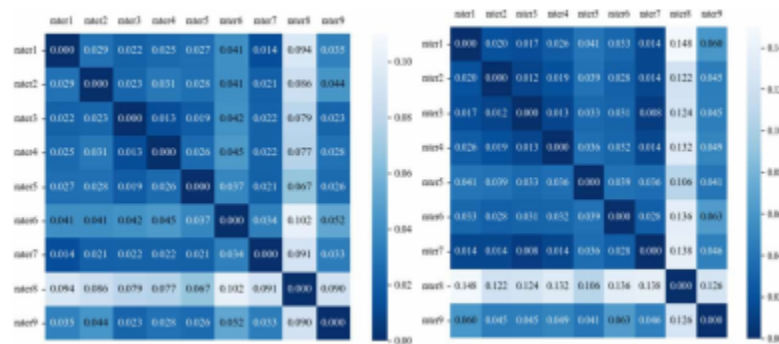
Spearman's coefficient	relevance	originality	clarity	specificity
relevance	1	0.778**	0.722**	0.670**
originality	0.778**	1	0.780**	0.772**
clarity	0.722**	0.780**	1	0.883**
specificity	0.670**	0.772**	0.883**	1
Kendall's coefficient	relevance	originality	clarity	specificity
relevance	1	0.713**	0.667**	0.605**
originality	0.713**	1	0.726**	0.711**
clarity	0.667**	0.726**	1	0.840**
specificity	0.605**	0.711**	0.840**	1

Note: ** Significance at the 0.01 level (two-tailed).

Table 7. Rank correlation coefficients between evaluation metrics (Ernie Bot)

Spearman's coefficient	relevance	originality	clarity	specificity
relevance	1	0.692**	0.695**	0.708**
originality	0.692**	1	0.740**	0.769**
clarity	0.695**	0.740**	1	0.876**
specificity	0.708**	0.769**	0.876**	1
Kendall's coefficient	relevance	originality	clarity	specificity
relevance	1	0.643**	0.646**	0.650**
originality	0.643**	1	0.686**	0.707**
clarity	0.646**	0.686**	1	0.846**
specificity	0.650**	0.707**	0.846**	1

Note: ** Significance at the 0.01 level (two-tailed).



(a) JS divergence heat map depicting the similarity of ratings between pairs of evaluators (ChatGPT).

(b) JS divergence heat map depicting the similarity of ratings between pairs of evaluators (Ernie Bot).

Figure 2. JS divergence heat map depicting the similarity of ratings between pairs of evaluators.

6 DISCUSSION

This work assessed the ability of ChatGPT and Ernie Bot to generate research priorities in the field of EECF, covering mechanisms, structural improvements, applications in cardiology, applications in neurology, and applications in other fields. Both models demonstrated significant potential in consistently generating relevant and clear research priorities, which could offer valuable new tools for EECF research. That said, both scored relatively low in specificity, possibly due to limitations in handling domain-specific knowledge, indicating a need for improvement in accuracy and precision. To enhance their performance, fine-tuning with domain-specific data and expert knowledge will likely be required. While both models lacked originality in their responses, relying heavily on learned information and language patterns, future research should focus on enhancing their creativity to generate more unique research questions in the EECF field.

Notably, this study also compared the performances of Ernie Bot and ChatGPT, two prominent language systems. Ernie Bot demonstrated a slight but definitive advantage in terms of relevance, possibly due to its more precise semantic understanding and higher matching with user needs. In terms of originality, ChatGPT scored slightly lower with more fluctuation, indicating some disagreement among evaluators regarding its ability to offer novel and unique perspectives. This variance might stem from differences in the models' performance across different contexts or from evaluators' subjective criteria, such as their acceptance of research priorities that challenge existing cognitive frameworks or their willingness to explore unknown areas of study. In contrast, Ernie Bot received more consistent recognition for its originality, likely due to its more flexible and innovative thinking patterns. Regarding clarity and specificity, both models performed equally well, demonstrating high levels of proficiency. This suggests that they excel in providing clear, understandable responses and specific, detailed explanations, which are equally important for large language models as users often expect answers that are both clear and specific to better understand and apply the provided information.

From the evaluators' perspective, most evaluators held similar views on the performance of the two models. However, in certain specific cases, such as Rater3 and Rater4, Ernie Bot received higher scores. Additionally, as compared to other raters, Rater8's scores were significantly lower and deviated more substantially, and exclusion of Rater8 increased the performance of both models.

In certain specific topics such as mechanisms, applications in neurology, and cardiovascular applications, Ernie Bot performed better whereas ChatGPT's performance slightly surpassed that of Ernie Bot in others, indicating that each model has its strengths and weaknesses in different domains and application scenarios. Therefore, future research could further explore how to integrate the strengths of both models to enhance the performance and effectiveness of large language models in practical applications.

While the study yields promising outcomes, there were some clear limitations. Firstly, the expert panels involved may not fully represent the broader research com-

munity which could have influenced evaluation outcomes. Secondly, the use of subjective ratings could have introduced potential bias and variability in assessing ChatGPT and Ernie Bot's performance. Lastly of note, the models may lack access to the latest biomedical literature, impacting question generation. If true, integrating domain-specific APIs with updated information could enhance research quality.

7 CONCLUSION

Overall, this assessment of ChatGPT and Ernie Bot as generators of research priorities for Enhanced External Counterpulsation (mechanisms, device improvements, applications in cardiovascular medicine, applications in neurology, and applications in other non-cardiovascular and non-neurological fields) produced some promising results. Both models have demonstrated the capacity to generate high-quality research priorities in these areas, indicating their potential value as tools to drive research not only in EECP but also in broader medical fields through streamlining the process of identifying crucial research priorities and thereby save considerable time and effort. While there is room for improvement in terms of specificity and originality, both models have shown a capability to produce diverse, relevant, and coherent research priorities, likely aiding advancements in EECP research. Each model has its strengths in various domains and application scenarios, and further exploration could focus on leveraging these strengths to enhance the overall effectiveness of large language models in practical settings. In conclusion, our findings suggest that ChatGPT and Ernie Bot are poised to become valuable assistants for researchers in the EECP field and likely other medical domains, offering new momentum for scientific progress.

References

1. Zhang Y, Enhanced external counterpulsation inhibits intimal hyperplasia by modifying shear stress responsive gene expression in hypercholesterolemic pigs. *Circulation*. 2007 Jul 31;116(5):526-34. doi: 10.1161/CIRCULATIONAHA.106.647248. Epub 2007 Jul 9. PMID: 17620513.
2. Lin S, Xiao-Ming W, Gui-Fu W. Expert consensus on the clinical application of enhanced external counterpulsation in elderly people (2019). *Aging Med (Milton)*. 2020 Mar 3;3(1):16-24. doi: 10.1002/agm2.12097. PMID: 32232188; PMCID: PMC7099759.
3. Xu L, Chen X, Cui M, et al. The improvement of the shear stress and oscillatory shear index of coronary arteries during Enhanced External Counterpulsation in patients with coronary heart disease. *PLoS One*. 2020 Mar 19;15(3):e0230144. doi: 10.1371/journal.pone.0230144. PMID: 32191730; PMCID: PMC7082042.
4. Kang, Y., Cai, Z., Tan, C. W., Huang, Q., & Liu, H. (2020). Natural language processing (NLP) in management research: A literature review. *Journal of Management Analytics*, 7(2), 139–172. <https://doi.org/10.1080/23270012.2020.1756939>
5. Lahat A, Shachar E, Avidan B, Glicksberg B, Klang E. Evaluating the Utility of a Large Language Model in Answering Common Patients' Gastrointestinal Health-Related Questions: Are We There Yet? *Diagnostics*. 2023; 13(11):1950.
6. Open AI. ChatGPT: Optimizing Language Models for Dialogue. Available at: <https://openai.com/blog/chatgpt/> (accessed 03/15/2024).
7. Ernie Bot: Baidu's Knowledge-Enhanced Large Language Model Built on Full AI Stack Technology. Available at: <https://yiyian.baidu.com/> (accessed 03/15/2024).
8. Gala D, Makaryus AN. The Utility of Language Models in Cardiology: A Narrative Review of the Benefits and Concerns of ChatGPT-4. *International Journal of Environmental Research and Public Health*. 2023; 20(15):6438. <https://doi.org/10.3390/ijerph20156438>
9. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the Feasibility of ChatGPT in Healthcare: An Analysis of Multiple Clinical and Research Scenarios. *J Med Syst*. 2023 Mar 4;47(1):33. doi: 10.1007/s10916-023-01925-4. PMID: 36869927; PMCID: PMC9985086.

10. George Pallivathukal R, Kyaw Soe HH, Donald PM, Samson RS, Hj Ismail AR. ChatGPT for Academic Purposes: Survey Among Undergraduate Healthcare Students in Malaysia. *Cureus*. 2024 Jan 27;16(1):e53032. doi: 10.7759/cureus.53032. PMID: 38410331; PMCID: PMC10895383.
11. Tang L, Sun Z, Idnay B, Nestor JG, Soroush A, Elias PA, Xu Z, Ding Y, Durrett G, Rousseau JF, Weng C, Peng Y. Evaluating large language models on medical evidence summarization. *NPJ Digit Med*. 2023 Aug 24;6(1):158. doi: 10.1038/s41746-023-00896-7. PMID: 37620423; PMCID: PMC10449915.
12. Balas M, Wadden JJ, Hébert PC, Mathison E, Warren MD, Seavilleklein V, Wyzynski D, Callahan A, Crawford SA, Arjmand P, Ing EB. Exploring the potential utility of AI large language models for medical ethics: an expert panel evaluation of GPT-4. *J Med Ethics*. 2024 Jan 23;50(2):90-96. doi: 10.1136/jme-2023-109549. PMID: 37945336.
13. Van Veen D, Van Uden C, Blankemeier L, Delbrouck JB, Aali A, Bluethgen C, Pareek A, Polacin M, Reis EP, Seehofnerová A, Rohatgi N, Hosamani P, Collins W, Ahuja N, Langlotz CP, Hom J, Gatidis S, Pauly J, Chaudhari AS. Adapted large language models can outperform medical experts in clinical text summarization. *Nat Med*. 2024 Apr;30(4):1134-1142. doi: 10.1038/s41591-024-02855-5. Epub 2024 Feb 27. PMID: 38413730.
14. Joseph, S.A., Chen, L., Trienes, J., Göke, H.L., Coers, M., Xu, W., Wallace, B.C., & Li, J.J. (2024). FactPICO: Factuality Evaluation for Plain Language Summarization of Medical Evidence. *ArXiv*, abs/2402.11456.
15. Pezeshkpour, P. Measuring and Modifying Factual Knowledge in Large Language Models. 2023 International Conference on Machine Learning and Applications (ICMLA), 831-838.
16. Guo, Y., Yang, Y., & Abbasi, A. . Auto-Debias: Debiasing Masked Language Models with Automated Biased Prompts. Annual Meeting of the Association for Computational Linguistics.(2022)
17. Yang DY, Wu GF. Vasculoprotective properties of enhanced external counterpulsation for coronary artery disease: beyond the hemodynamics. *Int J Cardiol*. 2013 Jun 5;166(1):38-43. doi: 10.1016/j.ijcard.2012.04.003. Epub 2012 May 4. PMID: 22560950.
18. Shea ML, Conti CR, Arora RR. An update on enhanced external counterpulsation. *Clin Cardiol*. 2005 Mar;28(3):115-8. doi: 10.1002/clc.4960280304. PMID: 15813616; PMCID: PMC6654750.
19. Gorcsan, J., et al. Improvement in left ventricular performance by enhanced external counterpulsation in patients with heart failure. *JOURNAL OF THE AMERICAN COLLEGE OF CARDIOLOGY*. 35(2). 655 AVENUE OF THE AMERICAS, NEW YORK, NY 10010 USA: ELSEVIER SCIENCE INC, 2000.
20. Éneeva MA, Kostenko EV, Razumov AN, Petrova LV, Bobyрева SN, Nesuk OM. [The enhanced external counterpulsation as a method of non-invasive auxiliary blood circulation used for the combined rehabilitative treatment of the patients surviving after ischemic stroke (a review)]. *Vopr Kurortol Fizioter Lech Fiz Kult*. 2015 May-Jun;92(3):45-52. Russian. doi: 10.17116/kurort2015345-52. PMID: 26285334.
21. Xiong L, Lin W, Han J, Chen X, Leung T, Soo Y, Wong KS. Enhancing cerebral perfusion with external counterpulsation after ischaemic stroke: how long does it last? *J Neurol Neurosurg Psychiatry*. 2016 May;87(5):531-6. doi: 10.1136/jnnp-2014-309842. Epub 2015 Apr 30. PMID: 25934015.
22. Sardina PD, Martin JS, Avery JC, Braith RW. Enhanced external counterpulsation (EECP) improves biomarkers of glycemic control in patients with non-insulin-dependent type II diabetes mellitus for up to 3 months following treatment. *Acta Diabetol*. 2016 Oct;53(5):745-52. doi: 10.1007/s00592-016-0866-9. Epub 2016 May 14. PMID: 27179825.
23. Froschermaier SE, Werner D, Leike S, Schneider M, Waltenberger J, Daniel WG, Wirth MP. Enhanced external counterpulsation as a new treatment modality for patients with erectile dysfunction. *Urol Int*. 1998;61(3):168-71. doi: 10.1159/000030315. PMID: 9933838.
24. Spearman C. The proof and measurement of association between two things. *Int J Epidemiol*. 2010 Oct;39(5):1137-50. doi: 10.1093/ije/dyq191. PMID: 21051364.
25. Kendall, M.G. (1938). A NEW MEASURE OF RANK CORRELATION. *Biometrika*. 30 (1–2): 81–89. doi:10.1093/biomet/30.1-2.81. JSTOR 2332226.