LEARNING LETTERS

Year 2024
Volume 3
Article 34
https://doi.org/10.59453/ll.v3.34

A comparative analysis of critical thinking skills between ChatGPT and college students

Shan Li[1], Fan Min[2], and Gurnam Kaur Sidhu[3]

[1] No. 2 Middle School of Yucheng, Ya'an, China

[2] Sichuan School of Foreign Languages, Sichuan University of Science and Engineering, China

[3] Faculty of Education, Languages, Psychology & Music, SEGi University, Malaysia

This study employed a questionnaire survey using the California Critical Thinking Skills Test (CCTST) on a sample of 240 Chinese university students. The results obtained were then compared with 240 questionnaire results created by ChatGPT in order to investigate variations in critical thinking skills across different dimensions. The study uncovered that, although ChatGPT had exceptional performance in inference and deductive reasoning, it fell behind university students in the areas of evaluation and inductive reasoning. Meanwhile, with respect to analysis, the capabilities of both college students and ChatGPT were not significantly different. These findings provide essential insights for the advancement of artificial intelligence and human progress.

Keywords: ChatGPT, college students, comparative study, critical thinking skills

Correspondence: Li Shan, sallyli9898@gmail.com

Recommended citation: Li, S., Min, F., & Sidhu, G. K. (2024). A comparative analysis of critical thinking skills between ChatGPT and college students. *Learning Letters*, 3, Article 34. https://doi.org/10.59453/ll.v3.34

Introduction

Artificial intelligence (AI) systems, particularly language models like ChatGPT, have exhibited remarkable capabilities in comprehending and generating human-like text (Baidoo-Anu & Owusu Ansah, 2023; Haleem et al., 2022). In addition, some studies have shown that ChatGPT possesses strong deductive reasoning and inference abilities (Kung et al., 2023; Maddigan & Susnjak, 2023; Sabzalieva & Valentini, 2023; Zhong et al., 2023). Reasoning is a fundamental component of critical thinking, playing a pivotal role in problem-solving and decision-making. Cultivating critical thinking skills in students is also highly important for human education. In the United States, fostering students' critical thinking skills is listed as a crucial educational objective (Roth, 2010).

Consequently, there is considerable interest in the extent to which AI can replicate human critical thinking across various dimensions. This research aims to explore the critical thinking skills of analysis, evaluation, inference, inductive reasoning, and deductive reasoning by comparing the performance of OpenAI's ChatGPT 3.5 with that of college students. The objective is to gain insights into ChatGPT's strengths and limitations when compared to human critical thinking skills.

Method

Participants

For this study, a total of 240 students were chosen at random from different departments of one of the largest universities in the southwestern area of China. In total, there were 89

males and 151 females, with ages ranging from 17 to 20 years old.

In accordance with the study's ethics protocol, all participants were informed that their data would be used exclusively for research purposes and that providing personal names was not required when completing the study questionnaire. Subsequently, consent was obtained from both individuals and the university authorities.

As a trailblazer in the field of large language models, ChatGPT 3.5 was one of the first to be publicly available. This open-source characteristic not only established it as a benchmark for future models, but also fostered collaboration and innovation within the research community. A new ChatGPT account was registered, and each of the 240 questionnaires was sequentially inputted into the system, recording the responses generated by ChatGPT 3.5.

Instrument

As critical thinking research has advanced, scholars have developed tests for assessing critical thinking. Such tests include Cornell Critical Thinking Test Levels X and Z (Ennis, 1993), California Critical Thinking Disposition Inventory (Facione et al., 1994), Ennis-Weir Critical Thinking Essay Test (Ennis & Weir, 1985) and Watson-Glaser Critical Thinking Appraisal (Watson, 1980). Regardless of the theoretical framework utilised in developing these assessments, they all share a fundamental feature: they measure basic critical thinking abilities that are not specific to any single subject.

The California Critical Thinking Skills Test (CCTST) focuses on cognitive abilities such as interpretation, analysis, evaluation, explanation and inference. Its primary purpose is to assess the critical thinking skills of college undergraduate students (P. A. Facione, 1990). In this study, students' critical thinking abilities were assessed using the Chinese version of the CCTST Form A (CCTST-A), which was created by Peter A. Facione and officially translated into Chinese by Luo Qing Xu. This questionnaire includes 34 items (P. A. Facione, 1990). Table1 displays the dimensions to which the 34 questions belong.

Table1. The sub-skills of CCTT and the items of sub-skills

Scale	Item
Analysis	1-9
Evaluation	10-13，24，26-34，
Inference	14-24
Deductive reasoning	1，2，5，6，11，19, 22，23，29
Inductive reasoning	10，11，20，21，24，25，26-28，30-34

The Chinese version of the CCTST demonstrates high reliability, as evidenced by a retest correlation of 0.63 ($p < 0.01$) over a one month period, as well as split-half correlations of 0.75 and 0.80 (both $p < 0.01$). Likewise, the Chinese CCTST has strong validity. Instruction in critical thinking has a substantial positive impact on test performance, demonstrating strong structural validity. In addition, there is a high correlation between the total scores on the Chinese version of the CCTST and students' GPAs (Grade Point Average) as well as their results on Raven's Standard Progressive Matrices test. This indicates a strong criterion validity (Luo, 2002). Each accurate response is awarded one point, resulting in scores ranging from 0 to 34. A higher score signifies greater proficiency in critical thinking skills.

Data collection and procedure

Because the questionnaire was large, complex, and time-consuming, printed paper versions were given to pupils. Students completed the CCTST within the confines of their classrooms, according to a strict time limit of 50 minutes. Afterwards, the author gathered the responses and uploaded them to Wenjuanxing, a site that provides similar functionalities to those found on Amazon Mechanical Turk (Wu et al., 2018). In order to analyse statistical data consistently and uncover differences and similarities between college students and ChatGPT, a total of 240 surveys were separately delivered by ChatGPT 3.5.

Multiple jobs were performed during the data purification operation conducted in OpenRefine. The tasks required converting 'N/A' strings into the numerical value '0'. The decision to transform 'N/A' into '0' was made in order to maintain data consistency and to effectively manage missing values during the analysis procedure. Abnormal numerical inputs were resolved by manually examining and correcting them to guarantee the precision and dependability of the data.

Data analysis

Dimension comparison

Prior to completing dimension comparisons, a normality test was conducted on the data, as depicted in Table 2. The sample size consisted of 480 instances.

Table 2. Shapiro-Wilk Normality Test result plot

Scales	Group	Schapiro-Wilk	Sig
Analysis	ChatGPT	0.941	0.000
	Students	0.942	0.000
Evaluation	ChatGPT	0.846	0.000
	Students	0.966	0.000
Inference	ChatGPT	0.956	0.000
	Students	0.962	0.000
Deductive reasoning	ChatGPT	0.930	0.000
	Students	0.970	0.000
Inductive reasoning	ChatGPT	0.854	0.000
	Students	0.968	0.000

Given the non-normal distribution of the data, the Shapiro-Wilk test was used. A significance level of $p < 0.05$ suggests a significant difference, indicating that the data for different dimensions in both groups did not adhere to a normal distribution. Therefore, dimension comparisons were conducted using independent sample non-parametric testing, as shown in Table 3.

Table 3. Mann-Whitney dimension comparison

Index	Group	Number of cases	Rank mean	Sig
Analysis	ChatGPT	240	248.68	0.183
	Students	240	232.32	
Evaluation	ChatGPT	240	218.29	0.000
	Students	240	262.71	
Inductive reasoning	ChatGPT	240	210.12	0.000
	Students	240	270.88	
Inference	ChatGPT	240	293.08	0.000
	Students	240	187.92	
Deductive reasoning	ChatGPT	240	296.61	0.000
	Students	240	184.39	

In the dimension of *analysis*, the value of p is greater than 0.05, which suggests that there is no substantial difference between ChatGPT and college students. Yet, in terms of *evaluation* and *inductive reasoning*, the statistical analysis ($p < 0.05$) indicates a noteworthy distinction, with students demonstrating a higher median rank compared to ChatGPT. This indicates that students surpass ChatGPT in these two areas. In the *inference* and *deductive reasoning* dimensions, the value of p is also less than 0.05, indicating a substantial distinction between the two groups. Specifically, ChatGPT has a higher median rank compared to the students. This suggests that ChatGPT outperforms university students in these two domains.

Overall, the analysis of the Shapiro-Wilk Test indicates notable discrepancies in the dimensions when comparing ChatGPT and college students. ChatGPT has suboptimal performance in *evaluation* and *inductive reasoning*, while displaying exceptional proficiency in *inference* and *deductive reasoning*. Both ChatGPT and the students demonstrated comparable levels of proficiency in *analysis* ($p > 0.05$), indicating that there is no statistically significant difference in this aspect.

Correlation between different dimensions

Correlation is a quantitative assessment of the degree of association between two or more variables. The examination of the associations between various aspects of the CCTST provided significant revelations.

In the Pearson correlation analysis, ** represents a p-value less than 0.01, and * represents a p-value less than 0.05. A p-value less than 0.05 indicates the presence of correlation, while a p-value less than 0.01 indicates a strong correlation (Hamdan et al., 2013). Tables 4 and 5 demonstrate that both ChatGPT and college students display high coefficients in different dimensions, indicating a strong correlation in critical thinking skills. This suggests that ChatGPT 3.5 and university students exhibit similar patterns of performance in these skills, possibly due to shared influences. This finding emphasises the complex connections between many dimensions. It is crucial to acknowledge that this outcome signifies associations, rather than causation. Hence, it should be noted that enhancing one dimension does not necessarily lead to the enhancement of another, particularly when comparing the capabilities of artificial intelligence to those of humans.

Table 4. College students: Pearson correlation analysis between dimensions

Scales	Analysis	Evaluation	Inference	Deductive reasoning	Inductive reasoning
Analysis	1				
Evaluation	.269**	1			
Inference	.167**	.285**	1		
Deductive reasoning	.403**	.489**	.753**	1	
Inductive reasoning	.321**	.821**	.493**	.418**	1

Note: ** indicates $p < 0.01$; * indicates $p < 0.05$

Table 5. ChatGPT: Pearson correlation analysis between dimensions

Scales	Analysis	Evaluation	Inference	Deductive reasoning	Inductive reasoning
Analysis	1				
Evaluation	.135*	1			
Inference	.183**	.200**	1		
Deductive reasoning	.309**	.447**	.772**	1	
Inductive reasoning	.213**	.842**	.398**	.326**	1

Note: ** indicates $p < 0.01$; * indicates $p < 0.05$

Conclusions

ChatGPT's performance in *evaluation* and *inductive reasoning* may not match that of university students, but it outperforms them in the *inference* and *deductive reasoning* dimensions. In the realm of *analysis*, there is no significant difference between ChatGPT and college students. As a large-scale language model, ChatGPT may fall short in certain aspects of higher-order thinking compared to humans (Deng et al., 2022; Guo et al., 2023b). However, this doesn't imply that humans will always maintain the upper hand in higher-order thinking abilities. It is essential to recognise that ChatGPT's *inference* and *deductive reasoning* capabilities have made significant advancements over time (H. Liu et al., 2023; Y. Liu et al., 2023). While this study provides valuable insights into the comparative critical thinking skills of humans and AI, several limitations should be acknowledged:

Sample representativeness: The study's participants were limited to Chinese university students. This specificity may affect the generalisability of the findings to other populations or educational contexts. Future research should include a more diverse sample to enhance the applicability of the results across different cultural and educational backgrounds.

AI version specificity: The study utilised a specific version of ChatGPT (version 3.5). The capabilities and performance of ChatGPT can vary between versions, so the findings may not be representative of other versions or future iterations of this AI tool. Future studies should consider evaluating multiple versions of AI models to account for their evolving capabilities.

Scope of CCTST: The California Critical Thinking Skills Test is a comprehensive tool, but it is only one of many instruments available to measure critical thinking. Future research could benefit from employing a variety of critical thinking assessments to capture different dimensions and aspects of this complex cognitive skill.

Funding

No funding was received for the conduct of this research.

Disclosure statement

The authors report no potential conflict of interest.

Disclosure of the use of AI-assisted technologies during writing

No AI-assisted technologies were used in the writing or editing of this manuscript.

About the authors

Shan Li holds a Master of Education degree from Sichuan University of Science & Engineering in China. She is currently a middle school English teacher with research interests in English teaching and modern educational technology.

Fan Min is a senior lecturer at the School of Foreign Languages at Sichuan University of Science & Engineering in China. She is currently pursuing a PhD in Education at SEGi University in Malaysia. Her research interests encompass innovative education, educational technology, English education, and cross-cultural studies.

Gurnam Kaur Sidhu is a PhD and professor at the Faculty of Education, Languages, Psychology & Music at SEGi University in Malaysia. Her areas of specialisation include teaching methods, teacher education, postgraduate supervision, TESL, and educational management.

References

Baidoo-Anu, D., & Ansah, L. O. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of AI*, *7*(1), 52–62. https://dx.doi.org/10.2139/ssrn.4337484

Deng, J., & Lin, Y. (2022). The benefits and challenges of ChatGPT: An overview. *Frontiers in Computing and Intelligent Systems*, *2*(2), 81–83. https://doi.org/10.54097/fcis.v2i2.4465

Ennis, R. H. (1993). Critical thinking assessment. *Theory into Practice*, *32*(3), 179–186.

Ennis, R. H., & Weir, E. E. (1985). *The Ennis-Weir critical thinking essay test: An instrument for teaching and testing*. Midwest Publications.

Facione, N. C., Facione, P. A., & Sanchez, C. A. (1994). Critical thinking disposition as a measure of competent clinical judgment: The development of the California Critical Thinking Disposition Inventory. *Journal of Nursing Education*, *33*(8), 345–350.

Facione, P. A. (1990). *The California Critical Thinking Skills Test – College level. Technical report #1. Experimental validation and content validity.* California Academic Press.

Fink, L. D. (2003). *A self-directed guide to designing courses for significant learning.* Jossey Bass.

Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., Yue, J., & Wu, Y. (2023). How close is ChatGPT to human experts? Comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*. https://doi.org/10.48550/arXiv.2301.07597

Haleem, A., Javaid, M., & Singh, R. P. (2022). An era of ChatGPT as a significant futuristic support tool: A study on features, abilities, and challenges. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, *2*(4), 100089. https://doi.org/10.1016/j.tbench.2023.100089

Halpern, D. F. (2014). *Critical thinking across the curriculum: A brief edition of thought & knowledge*. Routledge.

Hamdan, H., Issa, Z. M., Abu, N., & Jusoff, K. (2013). Purchasing decisions among Muslim consumers of processed halal food products. *Journal of Food Products Marketing*, *19*(1), 54–61. https://doi.org/10.1080/10454446.2013.724365

Kung, T. H., Cheatham, M., Medenilla, A., Sillos, C., De Leon, L., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J.,& Tseng, V. (2023). Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLoS Digital Health*, *2*(2), e0000198. https://doi.org/10.1371/journal.pdig.0000198

Liu, H., Ning, R., Teng, Z., Liu, J., Zhou, Q., & Zhang, Y. (2023). Evaluating the logical reasoning ability of ChatGPT and GPT-4. *arXiv preprint arXiv:2304.03439*. https://doi.org/10.48550/arXiv.2304.03439

Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., He, H., Li, A., He, M., Liu, Z., Wu, Z., , Zhao, L., Zhu, D., Li, X., Qiang, N., Shen, D., Liu, T., & Ge, B. (2023). Summary of ChatGPT-related research and perspective towards the future of large language models. *Meta-Radiology*, *1*(2), 100017. https://doi.org/10.1016/j.metrad.2023.100017

Lo, C. K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences*, *13*(4), 410.

Luo, Q. X. (2002). *批判性思维理论及其测评技术研究* [Doctoral dissertation, Nanjing Normal University]. http://gffiy28995338bdc041daso0ocuqnnvccf6q9q.fffb.tsg.suse.edu.cn/kcms2/article/abstract?v=6TwuVQQ8bfdAxQuSff2uWk2vVc63OiawK2AICBsad9z4g2nu6YDAXDDWr4mYOp_CR3lPBwxYdGpM744oHFrY2WU5JpjmwkStrOmq5GHl1dk_MB2w7KttOMwa0Kdva_Ju4QTIXPb78ycXEEXqpC_J3d0Kbl_IYv38b7qbDk6Y7py9UATo9mGT6EJY_nvcaql&uniplatform=NZKPT&language=CHS

Maddigan, P., & Susnjak, T. (2023). Chat2VIS: Generating data visualizations via natural language using ChatGPT, codex and GPT-3 large language models. *IEEE Access*, *11*, 45181–45193.

Rasul, T., Nair, S., Kalendra, D., Robin, M., de Oliveira Santini, F., Ladeira, W. J., Sun, M., Day, I., Rather, R. A., & Heathcote, L. (2023). The role of ChatGPT in higher education: Benefits, challenges, and future research directions. *Journal of Applied Learning and Teaching*, *6*(1), 41–56. https://doi.org/10.37074/jalt.2023.6.1.29

Roth, M. S. (2010). Beyond critical thinking. *The Chronicle of Higher Education*, *3*, 19.

Rudolph, J., Tan, S., & Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning and Teaching*, *6*(1), 342–363. https://doi.org/10.37074/jalt.2023.6.1.9

Sabzalieva, E., & Valentini, A. (2023). *ChatGPT and artificial intelligence in higher education: Quick start guide.* UNESCO.

Walker, S. E. (2003). Active learning strategies to promote critical thinking. *Journal of Athletic Training*, *38*(3), 263–267.

Watson, G., & Glaser, E. M. (1980). *Watson-Glaser critical thinking appraisal: Forms A and B; Manual*. Psychological Corporation.

Wu, S. J., Bai, X., & Fiske, S. T. (2018). Admired rich or resented rich? How two cultures vary in envy. *Journal of Cross-Cultural Psychology*, *49*(7), 1114–1143. https://doi.org/10.1177/0022022118774943

Zhai, X. (2022). *ChatGPT user experience: Implications for education.* SSRN. https://dx.doi.org/10.2139/ssrn.4312418

Zhong, Q., Ding, L., Liu, J., Du, B., & Tao, D. (2023). Can ChatGPT understand too? A comparative study on ChatGPT and fine-tuned BERT. *arXiv preprint arXiv:2302.10198*. https://doi.org/10.48550/arXiv.2302.10198