Article

Not peer-reviewed version

Assessing the Alignment of AI-Generated Evaluations with Human Judgment: A Case Study in an Entrepreneurship Pitch Competition

Shreya Rao *

Posted Date: 31 January 2024

doi: 10.20944/preprints202401.2211.v1

Keywords: AI evaluations; Human judgment; Pitch competition; ChatGPT; Entrepreneurial evaluations

Article

Assessing the Alignment of AI-Generated Evaluations with Human Judgment: A Case Study in an Entrepreneurship Pitch Competition

Shreya Rao

405 Meadowmont Lane Chapel Hill, NC 27517, USA; srao1@protonmail.com

Abstract: Amidst the rapid advancements in artificial intelligence (AI) applications, determining the alignment between AI evaluations and human judgment remains crucial, especially in sectors like finance, healthcare, and education. This study delves into a unique scenario—a teenage entrepreneur pitch competition, mirroring real-world challenges faced by these young innovators. The objective was to assess the congruence between evaluations by ChatGPT, an AI model, and human judges. The results indicate a remarkable alignment between AI and human evaluations across several criteria. The implications of this study are far-reaching, highlighting AI's potential in ensuring unbiased and reliable evaluations in pitch competitions and beyond. However, navigating the juncture of AI and human judgment mandates continued research and ethical considerations.

Keywords: AI evaluations; Human judgment; Pitch competition; ChatGPT; Entrepreneurial evaluations

Introduction

In the dynamic landscape of artificial intelligence (AI) and decision-making, the question of alignment between AI-generated assessments and human judgment has emerged as a pivotal concern.[1] This inquiry holds particular significance in domains where evaluations have far-reaching consequences, such as finance, healthcare, and education.

The increasing use of AI in decision-making processes has garnered substantial attention in the realm of cognitive computing. AI models, powered by deep learning and natural language processing, have demonstrated remarkable capabilities across various tasks, from image recognition to natural language understanding.[2] This technological value has led to growing interest in harnessing AI for decision support in diverse domains. Models like ChatGPT have been deployed to assess content quality, generate summaries, and even aid in medical diagnoses.[3] The ability of AI systems to execute these tasks accurately and consistently highlights their potential to redefine evaluation processes.

As AI-assisted evaluations become more prominent, ethical considerations must be factored. Issues related to transparency, fairness, and the potential for bias in AI-generated assessments are subjects of active research.[4,5] Ensuring equitable evaluations by AI systems is essential to maintain trust and prevent unintended repercussions.

Our study focuses on a unique context—a pitch competition for teenage entrepreneurs, which was scored by successful adult entrepreneurs based on standard judging criteria. We conducted a study to assess the alignment or lack, if any, between ChatGPT's evaluations and those of real human judges. The projects were evaluated based on a standard pitch deck application, reflecting a real-world scenario that teenage entrepreneurs face as they seek to turn their innovative ideas into reality.

This study illuminates the potential role of AI in the evaluation of young entrepreneurs' innovative ideas while considering the ethical and practical dimensions of such applications.

 doi:10.20944/preprints202401.2211.v1

Methods

Our study was initiated within the framework of a "Shark Tank"-inspired pitch competition exclusively for teenage entrepreneurs.

Each participating teenager submitted a standard pitch deck application, providing comprehensive insights into their respective projects. These applications included detailed information on vital aspects such as the business model, market potential, feasibility, and anticipated impact. The judging panel was composed of two seasoned adult entrepreneurs, distinguished by their accomplishments in the business realm.

The evaluation criteria used by the judges were aligned with industry standards commonly applied in entrepreneurial pitch competitions (see Table 1). These criteria encompassed a diverse range of attributes, including the clarity and structure of the pitch, the creativity and originality of the idea, market potential, feasibility and viability, business model, impact and social value, and scalability of the idea or product. The average of the scores assigned by the two judges for each criteria was used as the comparator.

To conduct AI evaluations, we leveraged ChatGPT, an advanced natural language processing model developed by OpenAI. ChatGPT was provided the pitch deck applications submitted by teenage entrepreneurs and the same standardized judging criteria employed by the human judges and asked to score the applications on each criteria.

The primary objective of our research was to assess the degree of alignment between the evaluations generated by ChatGPT and those derived from human judges. We executed paired t-tests for each evaluation criterion to ascertain whether statistically significant differences existed between the scores assigned by ChatGPT and the human judges. A p value of 0.05 was set to be statistically significant.

Results

The statistical analysis revealed no statistically significant differences between the assessments provided by real judges and ChatGPT for any of the evaluated criteria, including "Clarity and Structure of Pitch" ($p \approx 0.3333$), "Creativity and Originality of the Idea" ($p \approx 0.9511$), "Market Potential" ($p = 1.0000$), "Feasibility and Viability" ($p \approx 0.1449$), "Business Model" ($p = 1.0000$), "Impact and Social Value" ($p \approx 0.2162$), and "Scalability of Idea/Product" ($p \approx 0.1263$). In all cases, p-values exceeded the significance level, indicating a high level of alignment between human experts and ChatGPT's evaluations for the 13 projects. These results suggest that ChatGPT demonstrated consistency with human judgment within the dataset, highlighting its potential to provide reliable evaluations in various decision-making scenarios.

Discussion

Our study found no statistically significant differences between the evaluations provided by ChatGPT and those conducted by the seasoned adult judges across any of the evaluation criteria. This finding has substantial implications for the integration of AI into entrepreneurial pitch competitions and decision-making processes more broadly.

The findings offer several compelling implications for the role of AI in entrepreneurial evaluations and related decision-making contexts. The reliability can be especially valuable in scenarios where a large number of evaluations need to be conducted efficiently. Another role could be to assist judges by providing additional insights and perspectives, potentially leading to more informed and equitable decisions. The impartial nature of AI evaluations can contribute to the reduction of bias and subjectivity in the decision-making process, ensuring fair evaluations for all participants.

Building on these findings, future research in this domain can explore several avenues including continued development of AI models to enhance their capacity to align with human judgment across a broader spectrum of criteria and contexts. Investigating hybrid evaluation models that combine AI

 doi:10.20944/preprints202401.2211.v1

assessments with human judgment can provide a balanced and comprehensive approach to entrepreneurial evaluations.

Conclusions

In conclusion, this study contributes to our understanding of the role of AI in entrepreneurial evaluations. The absence of statistically significant differences between ChatGPT's assessments and those of human judges highlights the potential for AI to play a substantial and consistent role in evaluating innovative projects. As AI continues to evolve and mature, it offers the promise of enhancing decision support, reducing bias, and fostering fairness in entrepreneurial pitch competitions and beyond. However, ethical considerations and ongoing research remain important as we navigate this dynamic intersection of AI and human judgment.

Table 1.

Judging Parameter	
Clarity and Structure (10 points)	
Was the pitch clear and easy to follow?	
Did the participant present their ideas in a well-organized manner?	
Creativity and Originality (10 points)	
How innovative and unique is the product or idea?	
Does it stand out from other solutions in the market?	
Market Potential (10 points)	
Is there a clear target audience for the product/service?	
Did the participant demonstrate an understanding of the market demand?	
Feasibility and Viability (20 points)	
Is the product or idea realistic and practical?	
Were potential challenges and obstacles addressed?	
Business Model (10 points)	
Was there a clear and logical plan for generating revenue?	
Were the financial aspects well thought out?	
Impact and Social Value (20 points)	
Does the product or idea have a positive impact on society or the environment?	
Was this aspect effectively conveyed during the pitch?	
Scalability (10 points)	

 doi:10.20944/preprints202401.2211.v1

Does the product or idea have the potential to grow and expand?	
Were future plans outlined?	
Total Score: (out of 80)	

Table 2.

Human Judges													
CRITERIA	Project 1	Project 2	Project 3	Project 4	Project 5	Project 6	Project 7	Project 8	Project 9	Project 10	Project 11	Project 12	Project 13
Clarity and Structure of Pitch (10)	10	5	7	9	10	9	8	9	8	6	10	10	7
Creativity and Originality of the idea (10)	10	5	7	10	10	9	7	9	7	5	10	10	7
Market Potential (10)	8	2	7	8	9	8	6	8	7	5	8	8	5
Feasibility and Viability (20)	18	10	13	12	17	18	14	17	15	9	18	18	5
Business Model (10)	8	10	6	5	8	8	7	8	6	5	8	7	5
Impact and Social Value (20)	18	10	17	10	19	17	16	10	17	8	19	16	10
Scalability of idea/product (10)	8	1	5	7	9	5	5	5	6	6	9	8	4
Total Score	**80**	**43**	**62**	**61**	**82**	**74**	**63**	**66**	**66**	**44**	**82**	**77**	**43**
ChatGPT													
CRITERIA	Project 1	Project 2	Project 3	Project 4	Project 5	Project 6	Project 7	Project 8	Project 9	Project 10	Project 11	Project 12	Project 13
Clarity and Structure of Pitch (10)	7	8	9	9	8	9	9	7	9	8	9	6	7
Creativity and Originality of the idea (10)	6	6	8	8	9	8	9	9	8	7	10	8	6
Market Potential (10)	8	7	9	9	9	8	9	6	9	8	9	6	7
Feasibility and Viability (20)	14	8	16	17	17	15	17	12	8	15	9	12	8
Business Model (10)	7	7	9	8	8	7	8	7	8	8	8	5	5
Impact and Social Value (20)	12	5	18	15	18	12	18	18	9	10	10	10	6
Scalability of idea/product (10)	6	6	9	8	8	8	8	7	8	7	8	4	6
Total Score	**60**	**47**	**78**	**74**	**77**	**67**	**78**	**66**	**59**	**63**	**63**	**51**	**45**

Table 3.

Criteria	Mean	Standard Deviation	p-Value	Scoring Scale
Clarity and Structure of Pitch	7.69	1.38	~0.3333	Out of 10
Creativity and Originality	6.92	2.01	~0.9511	Out of 10
Market Potential	7.08	2.05	1.0000	Out of 10
Feasibility and Viability	13.38	3.43	~0.1449	Out of 20
Business Model	7.08	1.76	1.0000	Out of 10
Impact and Social Value	13.54	4.89	~0.2162	Out of 20
Scalability of Idea/Product	5.85	2.23	~0.1263	Out of 10

References

1. Xu, Y. et al. (2021). 'Artificial intelligence: A powerful paradigm for scientific research', The Innovation, 2(4).
2. Khurana, D., Koli, A., Khatter, K. et al. (2023). 'Natural language processing: state of the art, current trends and challenges', Multimed Tools Appl, 82, pp. 3713-3744.
3. Temsah, O. et al. (2023). 'Overview of Early ChatGPT's Presence in Medical Literature: Insights From a Hybrid Literature Review by ChatGPT and Human Experts', Cureus, 15(4).
4. Balasubramaniam, N. et al. (2023). 'Transparency and explainability of AI systems: From ethical guidelines to requirements', Information and Software Technology, 159.
5. Elbadawi, M (2024),The role of artificial intelligence in generating original scientific research,International Journal of Pharmaceutics,Volume 652, 2024,123741, ISSN 0378-5173.