

A Comparative Analysis of Human and A.I. Feedback on Business Idea Evaluation

Bart Meijer

University of Twente
P.O. Box 217, 7500AE Enschede
The Netherlands

Abstract:

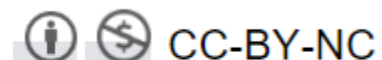
The field of artificial intelligence (A.I.) has grown significantly. Businesses and individuals are increasingly utilizing A.I. chatbots like ChatGPT to enhance efficiency and manageability, as finding novel business ideas is becoming more and more challenging due to market saturation. It is getting more interesting to use A.I. in business idea evaluation to give instantaneous and constructed feedback. This study examines business idea evaluations from an innovation tournament for novel business ideas. This thesis compares the evaluations made by humans and an A.I. chatbot. The evaluations on the applications were based on five criteria: Alignment, Excellence, Impact, Implementation, and Team. The study incorporated two distinct scenarios involving the A.I. chatbot. In one scenario, the chatbot received solely the evaluation form and the application, while in the other it was equipped with supplementary project information (prompting). This allowed for an assessment of the impact of prompting on the chatbot's evaluations. The results were then compared to human evaluations to assess the effectiveness and accuracy of the A.I. chatbot. This comparison was conducted using the T-test and the Man-Whitney U test to determine whether there were significant differences between the A.I. chatbot and human evaluations. According to these test there is a significant difference between humans and ChatGPT, with the exception of the evaluation from the more strict evaluating ChatGPT. This evaluation was not significantly different from the evaluation given by humans. It therefore suggests that given the right prompting ChatGPT is able to replicate the human evaluations. The findings of this study therefore have implications for the use of A.I. chatbots in evaluating business ideas and could streamline the evaluation process, making it more efficient and less reliant on human resources.

Graduation Committee members:

Michel Ehrenhard
Franziska Koefer
Tina Hormann

Keywords: ChatGPT, A.I., Feedback, Evaluation, Business idea evaluation

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.



1 Acknowledgments

I would like to thank my supervisors, Michel Ehrenhard, Franziska Koefer, and Tina Hormann, for all the support and great feedback I received on my work. The regular meetings really helped me make big steps in my thesis. I enjoyed working with them. Furthermore, I would like to thank my friends and family for the support they gave me through feedback and studying together. And lastly, I want to thank my fellow student, Nick Komonov, for providing me with a login for ChatGPT 4. Which made the use of the plugins possible.

2 Introduction

The field of artificial intelligence (AI) is rapidly expanding, and the emergence of AI chatbots like ChatGPT has enabled businesses and individuals to enhance efficiency and manageability in their lives. In the context of entrepreneurship, a crucial aspect of AI development lies in idea evaluation. Providing accurate and sufficient feedback during the idea development process can contribute to the creation of successful and innovative business ideas. The scope of this thesis is to find out how feedback given by A.I. compares to feedback given by humans. This is important because structured and fast feedback will greatly improve the idea generation process. It is getting more and more difficult to find novel ideas. This is due to the increasing number of new goods and services brought on by the speed of technological advancement, which has saturated the market and raised competition (Porter, 2008). Because many niches are already crowded with established businesses, this saturation makes it more challenging for entrepreneurs to identify untapped market opportunities (Kim & Mauborgne, 2009). More and more businesses are hosting tournaments to create novel ideas. With the rise of A.I. and the fast-growing interest in chatbots, especially ChatGPT, idea evaluation might become easier and easier with the use of A.I. chatbots. It would save lots of time and effort if feedback could be automated or partially taken over by A.I. As an A.I. chatbot can give feedback instantaneously and is less likely to make mistakes when given the right instructions (Haleem, Javaid, & Singh, 2023).

2.1 Research objective and research question

The objective of this research is to do a quantitative comparison between feedback provided by arti-

ficial intelligence(A.I.) chatbots and human experts concerning evaluating and improving business ideas. This research aims to identify the strengths and limitations of both A.I.-generated and human-generated feedback, while also analyzing the possible advantages when combining these two sources of input. The main goal is to provide scientific recommendations to those involved in the entrepreneurial ecosystem, including business consultants, investors, and entrepreneurs, on how to best combine A.I. and human ability to speed up the development and success of business ideas. To analyze this, the following research question is formulated and will be answered in this thesis:

How do evaluations of business ideas given by A.I. chatbots and human evaluators compare?

2.2 Relevance

2.2.1 Academic relevance

This thesis is academically relevant and this is demonstrated by the use of multiple theories such as the Power of Feedback theory (Hattie and Timperley, 2007) and the Social Information Processing Theory (SIP) (Walther, 1992), the proposed research design holds the potential to contribute to the existing body of knowledge in multiple fields, including human-computer interaction, entrepreneurship, and organizational behavior.

The Power of Feedback theory highlights the crucial role that feedback plays in shaping individuals' learning and performance (Hattie and Timperley, 2007). This theory is used in the research design to give useful feedback to business ideas that were generated through surveys and idea competitions. It has the potential to raise the quality of business ideas and promote entrepreneurial success.

The proposed research design is also academically significant because it compares the feedback given by human experts versus A.I. chatbots. The goal is to find out whether the feedback given by both is significantly comparable. This research extends the knowledge on this topic. It gives more knowledge on whether feedback given by A.I. can be interchanged already. The benefits and drawbacks of using A.I. systems for delivering feedback can help significantly affect the creation of more effective and efficient methods of reviewing business ideas as A.I. continues to develop.

In conclusion, the results of this study can help improve the knowledge of feedback processes, and technological adoption, as well as show how close technology comes to human evaluation scores.

2.2.2 Practical relevance

This research aims to find out whether the feedback provided by Artificial Intelligence (A.I.) is comparable to that given by human evaluators, and whether any observed similarities or differences are statistically significant. This gives insight into whether feedback given by A.I. can be perceived as similar when compared to feedback given by humans. Should the feedback provided by A.I. be found to be comparable to human feedback, it would suggest that A.I. has the potential to replicate human-like feedback in future applications and when A.I. will be able to give feedback that is as clear and as good as humans or even better. Giving good feedback will greatly improve the motivation and capabilities of people to improve their work to make it better suited to their needs. By using data from surveys and creating a framework for giving feedback, A.I. will be able to use these resources in future idea generation tournaments or in future studies to make giving feedback more efficient and instantaneous.

3 Literature review

This research will explore the comparison between feedback provided by artificial intelligence (A.I.) chatbots and human experts for evaluating and improving business ideas using a few main theories: the Power of Feedback, the Social Information Processing Theory (SIP), and Prompting. It will also give insight into the history of chatbots to be able to see how far knowledge and technology have come.

3.1 A.I.

In the growing field of artificial intelligence (A.I.), algorithms and computational models are created to help machines mimic human cognitive processes like learning, problem-solving, and decision-making (Russell, Norvig, & Davis, 2010). By delivering data-driven insights, automating tedious tasks, and improving decision-making abilities, artificial intelligence (A.I.) has the potential to fundamentally alter the way that business ideas are evaluated. Literature has identified numerous studies and applications of A.I. in assessing business ideas.

3.1.1 Natural language Processing(NLP)

In order to create algorithms and models for comprehending and producing human language, the multidisciplinary field of Natural Language Processing(NLP) combines expertise from computer science,

language study, and cognitive science to develop algorithms. NLP aims to develop intelligent systems capable of understanding, producing, and successfully interacting with humans through natural language.

NLP has a wide range of applications, including machine translation, sentiment analysis, information extraction, speech recognition, and conversational agents (Hirschberg & Manning, 2015). These applications have the potential to revolutionize various industries, such as healthcare, finance, education, and customer service, among others. "The ultimate effort is to make computer programs that can solve problems and achieve goals in the world as well as humans." (McCarthy, 2007) By implementing this knowledge Chatbots are being developed that have become incredibly great at natural language processing.

3.1.2 A.I. Chatbots

A.I. chatbots have been around for quite some time but in the last few decades, the capabilities and technologies behind these chatbots have drastically improved. One of the first times that computers were compared to humans was in a test created by Alan Turing .The aim of this test was to create a test in which individuals would engage in a conversation with a human and with a computer. They would then have to tell which they had the conversation with. If the individual could identify the computer as a human, the computer would have passed the Turing test (Turing, 1950). This test and research have caused researchers to do a lot more research and since the Turing test a lot more chatbots have been developed. The first one, ELIZA, was created by Joseph Weizenbaum in 1966 (Weizenbaum, 1966). ELIZA used simple rule-based systems and pattern matching to generate replies. This chatbot had limited conversational capabilities and frequently gave irrelevant or odd responses (Weizenbaum, 1966). Much later in the 21st century, the introduction of machine learning techniques, such as Artificial Neural Networks (ANN), allowing for more advanced chatbots like ALICE (Wallace, 2009). ALICE used a database of pre-defined responses and a heuristic algorithm to select the most appropriate response based on the user's input. Despite these improvements, these chatbots still lacked the ability to understand complex language patterns and generate contextually appropriate responses. With the development of GPT, OpenAI set the next step in NLP and the development of Chatbots.

3.1.3 ChatGPT

An important step forward for A.I. chatbots was made with the creation of Generative Pre-Trained Transformer(GPT) by OpenAI (Radford & Narasimhan, 2018). The foundation of GPT is the Transformer design, which uses self-attention methods to better capture long-range dependencies in text (Vaswani et al., 2017). This is the foundation of the GPT. GPT models are better able to produce responses that are relevant and logical since they have already been pre-trained on huge text data sets.

GPT-2 (Radford et al., 2019) displayed outstanding performance in challenges requiring natural language generation and understanding. The model's capacity to produce text that resembles human speech raised fears about potential abuse, such as the production of spam or fake news (Radford et al., 2019). Despite these fears, GPT-2 has been widely used for a variety of purposes, including summarizing, translating, and using chatbots.

Shortly after GPT-2 came GPT-3 (Brown et al., 2020), GPT-3 has further advanced the capabilities of A.I. chatbots. With 175 billion parameters, GPT-3 can generate well-structured, contextually appropriate, and creative responses, even in complex and ambiguous situations (Brown et al., 2020). GPT-3's performance has been compared to human-level comprehension in some tasks (Brown et al., 2020).

In the year 2023 came GPT-4, which improved on multiple levels, it can take image and text inputs and output text. GPT-4 performs at a level comparable to humans on a variety of professional and academic criteria, but being less effective than humans in many real-world circumstances (OpenAI, 2023).

3.2 Prompting

Prompting is the process of designing, utilizing and optimizing effective prompts to make sure that the chatbot does exactly what is intended (Ekin, 2023). Large Language Models (LLMs) can execute a range of tasks better thanks to the prompt engineering technique, including generating text, translating languages and producing other types of creative content. Short texts called prompts are used to direct LLM's responses. This can be really useful to make ChatGPT respond in the direction you want it to. Prompting will be used in this thesis to make sure that ChatGPT understands what it is asked to do in a more detailed way and does what it is supposed to do.

3.3 The Power of Feedback

It is well known that feedback has an important effect on the way individuals perform and their perception of their personal skills (Hattie & Timperley, 2007). Feedback plays a critical role in the evaluation and improvement of business ideas for both entrepreneurs and investors. The structure, application, and efficacy of feedback produced by A.I. chatbots and human experts will be examined using the Power of Feedback theory. The goal is to recognize the benefits and drawbacks of the feedback offered by these two sources and decide how to best combine it to have the biggest impact on the growth of business ideas. Feedback can also play a significant role in idea generation and innovation contests by providing constructive input and guidance to participants (Terwiesch & Ulrich, 2009). The qualities of ideas generated in such contests are influenced by the feedback process, as well as by the number of participants involved (Boud & Molloy, 2012). By assessing and giving feedback on the ideas generated, individuals can improve their understanding and refine their ideas, leading to higher-quality ideas (Nicol & Macfarlane, 2006). By having an A.I. chatbot that will be able to replicate the quality of feedback that humans can give, feedback during idea tournaments could be given while answering the questions for the tournament. This will lead to people getting direct feedback, which will help them improve directly instead of having to wait for a human to evaluate their input.

3.4 Social Information Processing Theory (SIP)

In Computer-Mediated Communication (CMC), the Social Information Processing Theory focuses on how people form opinions and make decisions based on minimal information. This theory helps in comprehending the distinctions between feedback generated by humans and A.I. when evaluating business ideas. It explains why people make assumptions through written messages and that given enough time people can feel the same intimacy as they do with humans in face-to-face meetings (Walther, 1992). To ensure that this interaction between humans and the A.I. is established in a way that promotes trust, understanding, and effective collaboration, guidelines for human-A.I. interaction can be established (Amershi et al., 2019). These guidelines can help in the development of A.I.-driven feedback systems that are both efficient and well-received by users. If A.I. is as proficient at giving feedback as humans then it is essential

that the people wanting to get feedback also accept that the A.I. is helping them.

4 Methodology

4.1 Digital platform for regional development

In this study, Data was used from an EU-funded research project that aims to promote the development of rural areas in Europe through digital innovation

The digital platform for regional development Project is an EU-funded research project that aims to promote the development of rural areas in Europe through digital innovation. The project develops a digital platform for services. The objective is to enhance the digital skills and literacy of rural citizens, providing them with access to digital tools and technologies, and fostering a culture of digital innovation in rural communities.

4.2 Idea Tournament

An Idea generation tournament has been held for the digital platform for regional development project. In this tournament companies described which rural need they were going to solve with the Digital platform for regional development and why this is significant for their area of interest and their rural stakeholders. From these entries, 15 entries were eligible and were evaluated by 2 evaluators for each of the eligible entries. This evaluation was done based on 5 criteria which were each given a score from 1 to 5. In this thesis these scores will be compared to the scores that A.I. gives to the same entries using the same evaluation sheet.

4.3 Research design

The innovation tournament will serve as the basis for the comparison between human and A.I. feedback. The evaluation paper used by humans to evaluate the 15 applications for the innovation tournament will be used to make A.I. evaluate the applications as well. Then the following process will follow to see if humans and A.I. can be seen as significantly different. Two different instances of A.I. will be used. One instance will be where the A.I. chatbot will not be given any other information than the evaluation form and the application, whereafter it will be asked to evaluate the application according to the evaluation form. The second instance will be where the chatbot is given proper prompting before getting the same information. This prompting will entail basic

information about the tournament and what the expectations of the chatbot will be. So the design of the research will be the following:

- All the applications will be anonymized to make sure that they can be used by Chatbots without giving away confidential details about people or companies.
- The most important information from the evaluation document will be clearly described so it can be easily comprehended by the A.I. chatbot.
- The first instance where no prior information is given will be put into the chatbot. Then the chatbot will evaluate the application according to the evaluation framework.
- The second instance where the prior information is given, will be put into the chatbot. Then the chatbot will evaluate the application according to the evaluation framework.
- The last instance where the chatbot is asked to give a more strict evaluation is put into the chatbot.
- Scores from the two human evaluators will be averaged.
- All the scores received from chatbot will be averaged as well.
- These scores received from the humans and the A.I. will be sorted by the 5 criteria.
- The scores given by the experts and A.I. will be compared on each evaluation criterion for each application.
- Determine agreement: Calculate the comparability between the humans and A.I. using statistical methods, such as the T-test and the Mann-Whitney U test.
- Check to see whether differences between the different groups are significant.

By following this process, the evaluation scores of the groups can be compared. The data that will be used will be that of the idea generation tournament of the digital platform for regional development.

4.4 Design Framework

To design the framework that is used for evaluating the business ideas, an evaluation form of the idea tournament of the digital platform for regional development will be used also to be able to compare the previous entries which have been already evaluated. Therefore the ideas will be evaluated on 5 criteria to make sure they fit within the digital platform for regional development platform these will be the following:

- Alignment
- Excellence
- Impact
- Implementation
- Team

These criteria were used because they have been used in the previously mentioned tournament. By using the same criteria it is easier to compare the human and ChatGPT evaluations because the human evaluations have already been evaluated according to a framework. These 5 criteria will be evaluated from 1 to 5 this and the meaning of the criteria will be further explained in the next subsection.

4.5 Prompting

By making the chatbot more familiar with its task it will be able to analyze the applications better and therefore have a closer resemblance to the information that the human evaluators have. As a result, the feedback provided by the chatbot becomes more closely aligned with that of the humans. The open calls objective gives good insight into what the actual project entails and what is expected from the applications. ChatGPT will be given the following prompts to ensure that it knows what the project is and therefore have a better ability to assess the applications:

"You are an expert analyzer for an innovation tournament for the digital platform for regional development, The project is an open call of which the goal is: "co-develop and implement a digital solution that delivers multiple services to rural citizens while creating opportunities for economic growth and quality of life improvements". More information that is essential for grading the application and get a better understanding of the open call can be found in the following file: <https://tinyurl.com/4xzns8zn>. please look at this file before grading the application. Your goal is to evaluate to the best of your abilities the

quality of applications for this open call. You will do this according to an evaluation framework containing 5 criteria which are ranked using a Likert scale from 1 to 5. The evaluation framework will first describe the criteria and what they mean and afterward it will explain how the rankings from 1 to 5 are described:

- Alignment
Extent to which the proposed business idea aligns with the digital platform for regional development project in general, and more specifically with the digital platform for regional development Open Call' objectives and requirements, regional marketplaces and strategic arenas and the extent to which the service proposed solves a rural need identified and addressing it is significant for the relevant stakeholders, providing specific value and being different from competition.
- Excellence
Extent to which applications demonstrate technical excellence, high level of innovativeness and ability to go beyond the state-of-the-art. The applicant should explain the technology behind the project and how it will address the challenges and TRL selected with its main outcomes. The applicant should provide information about the level of innovation within their market, degree of differentiation that this project will bring and to what extent the solution will go beyond the State of Art. The applicants should demonstrate how they plan the platform development and what is their ambition to do it with the digital platform for regional development support.
- Impact
Extent that the application is ambitious to improve quality of life in rural areas, indicating expected economic, technological, and societal impact on the Mirror Region. Strength of the marketing strategy and plan to reach customers. Credibility of the applicants' strategy to exploit the platform beyond the digital platform for regional development open call period and that there is a real interest in following up commercially.
- Implementation
Extent that the work plan is complete and coherent to ensure effective implementation. The soundness of the approach.
- Team Appropriateness of the team to meet the the digital platform for regional development

challenge. Technical capacity, experience, excellence, and quality of the team.

These 5 criteria need to be evaluated from 1 to 5, These will have the following values:

1. The application fails to address the issue under examination or provides insufficient information.
2. The issue is addressed in an unsatisfactory manner.
3. The application addresses the issue, but there are significant weaknesses.
4. The application addresses the issue well, but there is clear room for improvement.
5. The application successfully addresses all relevant aspects of the issue in question and if there are any shortcomings, they are minor.

please evaluate it by reviewing all the criterions and rating them according to what is mentioned above. You can use half points This is the document: 'Link to document'. make sure to take everything from the document into account and analyze the criterions one by one.

4.6 Statistical tests

4.6.1 T-Test

In this study, the T-test is used to determine whether the mean of the evaluation scores of two different variables is significantly different. The T-test is a statistical test that is used to compare the means of two groups. It is a parametric test, which means that it assumes that the data is normally distributed. The T-test can be used to compare two independent groups or two dependent groups (Cohen, 1988). In the case of this thesis, independent groups will be used. As there is no dependency between the human and A.I. evaluations. T-tests can only be performed if it conforms to certain assumptions. These assumptions are:

- It needs to have an ordinal or continuous scale
- It needs to be a random sample
- It needs to have a normal distribution
- A reasonably large sample size is used
- It needs homogeneity of variance

4.6.2 Non-parametric tests

When the sample distribution is not normally distributed or if there is no homogeneity of variance, non-parametric tests are a useful tool to be able to compare variables. They are statistical methods that do not assume a specific distribution of the data.

One of the most common non-parametric tests for comparing two independent samples is the Mann-Whitney U test, also known as the Wilcoxon rank-sum test. This test compares the medians of two groups and is often used with ordinal data or when the data are not normally distributed (Mann & Whitney, 1947). In this study quite a large portion of the data is not normally distributed and therefore the use of the Man-Whitney U test is useful.

The Mann-Whitney U test works by ranking all the observations from both groups together, then comparing the sum of ranks for each group. The null hypothesis is that the distributions of both groups are equal. If the test statistic is significantly different from what would be expected under the null hypothesis, this suggests a significant difference between the groups (Mann & Whitney, 1947).

This test needs to have certain assumptions met:

- Independence
The two samples that are being compared should be independent of each other. This means that the observations in one sample do not affect the observations in the other sample, and the other way around.
- The Mann-Whitney U test requires that the data be ordinal or continuous. This means that the data can be ranked, in the case of this thesis a Likert scale is used. This form of ranking falls under the category of ordinal data.
- The shape of the distribution
Unlike many parametric tests, the Mann-Whitney U test does not assume that the data follows a specific distribution. It does however assume that the shape of the distribution is the same for both groups. This means that if one group's ratings are skewed to the right, the other group's ratings should also be skewed to the right to the same degree (*Assumptions of the Mann-Whitney U test | Laerd Statistics, n.d.*).

5 Data collection

For the collection of the data, all the applications were evaluated according to the framework given above where the non-prompted ChatGPT was not

given the additional information that was provided before explaining the criteria. ChatGPT 4, equipped with plugin capabilities, utilized the "Ask Your PDF" plugin for this research. This plugin enables the upload of a PDF file, which can then be accessed and analyzed by ChatGPT. The use of this plugin was necessitated by the word input capacity limitations of ChatGPT when plain text was used.

Following the upload of the document, the evaluation framework as aforementioned was presented to ChatGPT, accompanied by the request to evaluate the uploaded document. ChatGPT proceeded to scan the entire document and evaluate it based on the established framework a few examples can be seen in Appendix A. Sometimes ChatGPT would briefly mention all the criteria after which it would give an evaluation score between 1 and 5 as seen in fig. 21. In other occasions ChatGPT would scan the document per 1 criteria and then thoroughly explain why it gave a specific grade as can be seen in Appendix A in fig. 10.

The prompted ChatGPT was first introduced to an extra document, it was asked to remember and use this document in the evaluation as can be seen in Appendix A in fig. 6. The answer to this can be found in fig. 7. The prompted Chatgpt again gave different answers a few examples of these answers can be found in Appendix 1 in fig. 8 and in figs. 16 to 20.

The stricter ChatGPT was asked after evaluating to evaluate it again but to be more strict as can be seen in Appendix A in figs. 9 and 11 to 15.

It was observed that the evaluations provided by the chatbot exhibited slight variations even when the same prompt was used. To account for this variability and enhance reliability, each application was evaluated three times by the chatbot, and an average score was calculated.

The resulting averages were placed into Excel tables to create a clear overview of all evaluation scores. Using this methodology, a total of 225 evaluation scores were generated by the non-prompted chatbot, and 225 average evaluation scores were produced by the prompted chatbot for each criterion.

In contrast, the human evaluations consisted of scores from two evaluators, resulting in a total of 150 evaluation scores. The average scores for each criterion from both the human evaluators and ChatGPT were then imported into SPSS for statistical analysis. An example of the layout of the SPSS data can be found in the Appendix in fig. 21

6 Results

The scores from the evaluation given by humans and A.I. then needs to be compared to see whether the values of these different groups are significantly different. The test needed for comparing two means is the T-test, as aforementioned this test requires assumptions to be met in order for it to be reliable. Therefore the assumptions need to be tested for all the comparisons between the groups. To check the normality of the samples a Shapiro-Wilk test is done to determine whether the samples are normally distributed.

6.1 Humans and non-prompted A.I.

First, the data will be checked for normality using the Shapiro-Wilk Test. These results are displayed in table 1

Criterion	statistic	df	sig.
1	.949	30	.003
2	.879	30	.160
3	.856	30	<.001
4	.863	30	.001
5	.858	30	<.001

Table 1: Shapiro-Wilk Test Humans and non-prompted ChatGPT

Every one of the samples can not be considered a normally distributed sample except for criterion 2. So for criterion 2, a T-test will be performed whereas for the rest a Man-Whitney U test will be performed and the assumptions need to be tested for all the criteria. Firstly, the assumptions needed for the Man-Whitney U test will be tested which have been mentioned in part 4.6.2. All the samples are independent samples, they were taken at different times and are not correlated. Secondly, the data of all the samples is a Likert scale which is an ordinal scale. The last assumption looks at the shape of the distribution of both groups and these need to be similar. In the following figure, the shape of the distribution for criterions 1,3,4, and 5 can be seen.

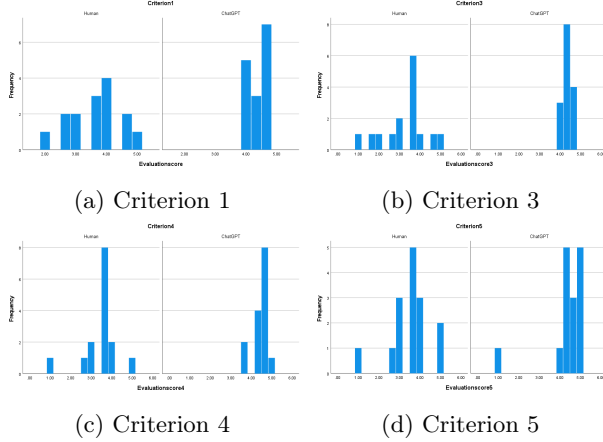


Figure 1: Distributions of Evaluation Scores for Man-Whitney U test Humans & ChatGPT

The distributions look like they are of approximately the same skewness. So all the assumptions are met and the Man-Whitney U test can be executed. For Criterion 2 the homogeneity of variances can be assumed by using the Levenes statistic. There was a homogeneity of variances for the evaluation scores for humans and ChatGPT, as assessed by Levene’s test for equality of variances ($p = 0.132$). Therefore the T-test can be used for criterion 2.

Criterion 1 A Man Whitney U test was performed to evaluate whether the scores given by humans for criterion 1 are significantly different from the scores given by the A.I. chatbot. ChatGPT had significantly higher evaluation scores than the human evaluators, $z = -2.990$, $p = 0.003$.

Criterion 2 The 15 averages of the evaluation scores given by humans ($M=3.07, SD=0.68$), compared to the 15 averages of the evaluation scores given by ChatGPT ($M=4.21, SD=0.38$) are significantly different, $t(28)=-5.721$, $p<0.01$.

Criterion 3 A Man Whitney U test was performed to evaluate whether the scores given by humans for criterion 3 are significantly different from the scores given by the A.I. chatbot. ChatGPT had significantly higher evaluation scores than the human evaluators, $z = -3.471$, $p < 0.001$.

Criterion 4 A Man Whitney U test was performed to evaluate whether the scores given by humans for criterion 4 are significantly different from the scores given by the A.I. chatbot. ChatGPT had significantly higher evaluation scores than the human evaluators, $z = -3.781$, $p < 0.001$.

Criterion 5 A Man Whitney U test was performed to evaluate whether the scores given by humans for criterion 5 are significantly different from the scores given by the A.I. chatbot. ChatGPT had significantly

higher evaluation scores than the human evaluators, $z = -2.986$, $p = 0.002$.

6.2 Humans and prompted GPT

First, the data will be checked for normality using the Shapiro-Wilk Test again. These results are displayed in table 2 and therefore the T-test cannot be used and the Man-Whitney U test will be used to compare the samples. The results can be found in fig. 2

Criterion	statistic	df	sig.
1	.843	30	<.001
2	.878	30	.002
3	.830	30	<.001
4	.862	30	.001
5	.900	30	.008

Table 2: Shapiro-Wilk Test Humans and Prompted ChatGPT

For this test, the assumptions have to be tested again. All the samples are independent samples, they were taken at different times and are not correlated. Secondly, the data of all the samples is a Likert scale which is an ordinal scale. The last assumption looks at the shape of the distribution of both groups and these need to be similar. In fig. 2, the shape of the distribution for the criteria can be seen.

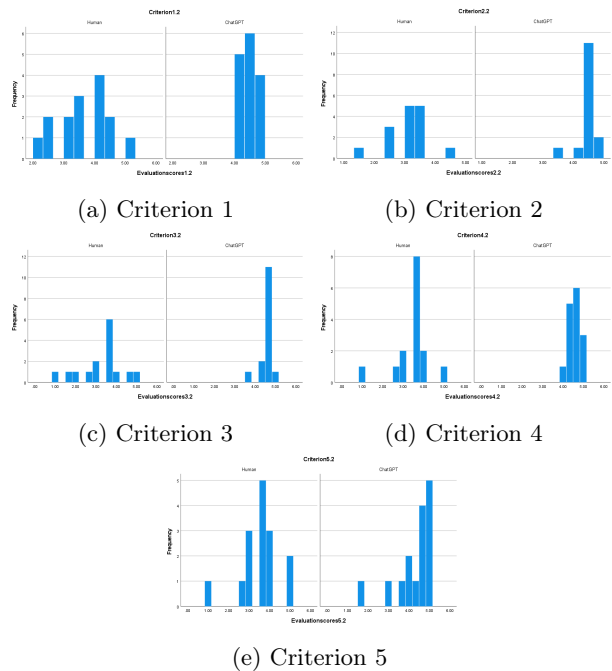


Figure 2: Distributions of Evaluation Scores for Man-Whitney U test Humans and ChatGPT prompted

The distributions look like they are of approximately the same skewness. Therefore all the assumptions of the tests are met.

Criterion 1 A Man Whitney U test was performed to evaluate whether the scores given by humans for criterion 1 are significantly different from the scores given by the A.I. chatbot. ChatGPT had significantly higher evaluation scores than the human evaluators, $z = -3.346$, $p < 0.001$.

Criterion 2 A Man Whitney U test was performed to evaluate whether the scores given by humans for criterion 2 are significantly different from the scores given by the A.I. chatbot. ChatGPT had significantly higher evaluation scores than the human evaluators, $z = -4.268$, $p < 0.001$.

Criterion 3 A Man Whitney U test was performed to evaluate whether the scores given by humans for criterion 3 are significantly different from the scores given by the A.I. chatbot. ChatGPT had significantly higher evaluation scores than the human evaluators, $z = -3.788$, $p < 0.001$.

Criterion 4 A Man Whitney U test was performed to evaluate whether the scores given by humans for criterion 4 are significantly different from the scores given by the A.I. chatbot. ChatGPT had significantly higher evaluation scores than the human evaluators, $z = -4.059$, $p < 0.001$.

Criterion 5 A Man Whitney U test was performed to evaluate whether the scores given by humans for criterion 5 are significantly different from the scores given by the A.I. chatbot. ChatGPT had significantly higher evaluation scores than the human evaluators, $z = -2.262$, $p = 0.023$.

6.3 ChatGPT prompted vs not prompted

First, the data will be checked for normality using the Shapiro-Wilk Test. These results are displayed in table 3 when looking at the results for the Shapiro-Wilk test it can be seen that all the samples are not normal and therefore the T-test cannot be used and the Man-Whitney U test will be used to compare the samples.

Criterion	statistic	df	sig.
1	.887	30	.004
2	.922	30	.029
3	.871	30	.002
4	.853	30	<.001
5	.713	30	<.001

Table 3: Shapiro-Wilk Test Humans and Prompted ChatGPT

For this test, the assumptions have to be tested again. All the samples are independent samples, they were taken at different times and are not correlated. Secondly, the data of all the samples is a Likert scale which is an ordinal scale. The last assumption looks at the shape of the distribution of both groups and these need to be similar. In fig. 3, the distribution for the sample can be found.

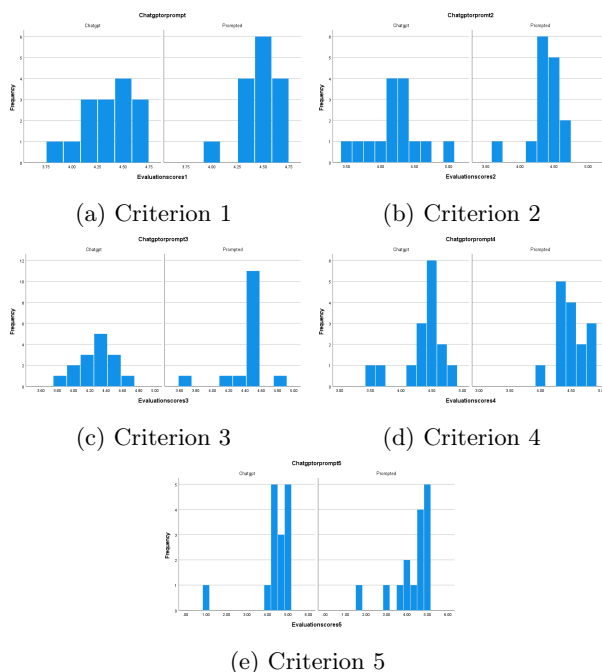


Figure 3: Distributions of Evaluation Scores for Man-Whitney U test of ChatGPT prompted & ChatGPT not prompted

The distributions look like they are of approximately the same skewness. Therefore all the assumptions of the tests are met.

Criterion 1 A Man Whitney U test was performed to evaluate whether the scores given by the prompted A.I. chatbot for criterion 1 are significantly different from the scores given by the A.I. chatbot. Because $p > 0.05$ we have to accept the null hypothesis and therefore it can be concluded that for this criterion

there is no significant difference between ChatGPT and ChatGPT prompted, $z = -1.242$, $p = 0.214$

Criterion 2 A Man Whitney U test was performed to evaluate whether the scores given by the prompted A.I. chatbot for criterion 1 are significantly different from the scores given by the A.I. chatbot. Because $p > 0.05$ we have to accept the null hypothesis and therefore it can be concluded that for this criterion there is no significant difference between ChatGPT and ChatGPT prompted, $z = -1.894$, $p = 0.058$.

Criterion 3 A Man Whitney U test was performed to evaluate whether the scores given by the prompted A.I. chatbot for criterion 3 are significantly different from the scores given by the A.I. chatbot. ChatGPT prompted had significantly higher evaluation scores than the human evaluators, $z = -2.332$, $p = 0.02$.

Criterion 4 A Man Whitney U test was performed to evaluate whether the scores given by the prompted A.I. chatbot for criterion 4 are significantly different from the scores given by the A.I. chatbot. Because $p > 0.05$ we have to accept the null hypothesis and therefore it can be concluded that for this criterion there is no significant difference between ChatGPT and ChatGPT prompted, $z = -0.749$, $p = 0.454$

Criterion 5 A Man Whitney U test was performed to evaluate whether the scores given by the prompted A.I. chatbot for criterion 5 are significantly different from the scores given by the A.I. chatbot. Because $p > 0.05$ we have to accept the null hypothesis and therefore it can be concluded that for this criterion there is no significant difference between ChatGPT and ChatGPT prompted, $z = -0.105$, $p = .917$

6.4 Stricter ChatGPT

Because of the big difference in the mean scores given by humans and the chatbots as can be seen in table 4, it can be insinuated that the chatbot should evaluate the applications more strictly, therefore the chatbot was asked to evaluate the application more strictly twice. This is for the reason to see whether a more strict evaluation is in fact significantly comparable to humans. The chatbot will again scan the document and apply a stricter evaluation method.

	Mean	Variance	stdev
Human	3.3267	.787	.8872
ChatGPT	4.3044	.256	.50625
ChatGPT Prompted	4.4044	.211	.45945

Table 4: Descriptives for Human, ChatGPT and ChatGPT prompted

This data will have to be checked for normality

again using the Shapiro-Wilk Test. These results are displayed in table 5

Criterion	statistic	df	sig.
1	.946	30	.135
2	.888	30	.004
3	.880	30	.003
4	.795	30	<.001
5	.914	30	.018

Table 5: Shapiro-Wilk Test Humans and Stricter ChatGPT

From this test, it can be concluded that only criterion 1 can be seen as a normal distribution and therefore this criterion will be checked for the assumptions of a T-test whereas for the rest a Man-Whitney U test will be performed and the assumptions need to be tested for all the criteria. Firstly, the assumptions needed for the Man-Whitney U test will be tested which have been mentioned in part 4.6.2. All the samples are independent samples, they were taken at different times and are not correlated. Secondly, the data of all the samples is a Likert scale which is an ordinal scale. The last assumption looks at the shape of the distribution of both groups and these need to be similar. In fig. 4, the shape of the distribution for criterions 2,3,4, and 5 can be seen.

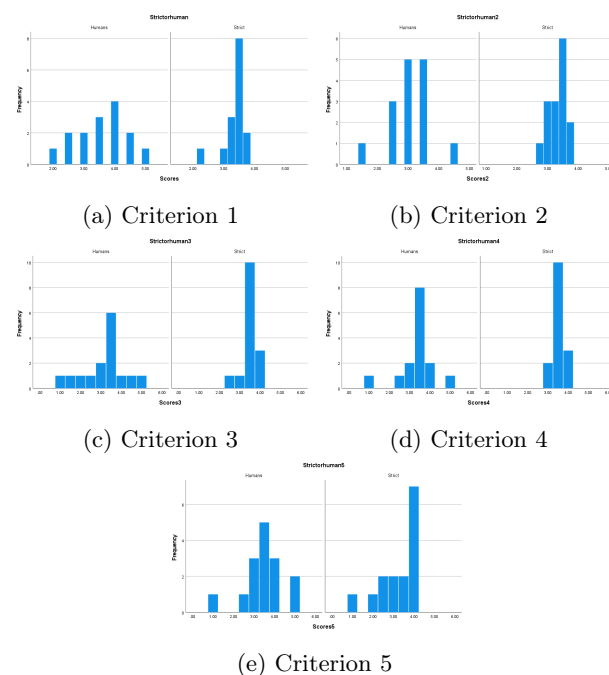


Figure 4: Distributions of Evaluation Scores for Man-Whitney U test Humans & Stricter ChatGPT

The skewness of the distributions is approximately the same and therefore this assumption is met as well.

For Criterion 1 the homogeneity of variances can not be assumed by using Levene’s statistic. There was not a homogeneity of variances for the evaluation scores for humans and the strict ChatGPT was assessed by Levene’s test for equality of variances ($p = 0.005$). Therefore the T-test cannot be used for criterion 1. So for all the criteria, the Man-Whitney U test will be executed.

Criterion 1 A Man Whitney U test was performed to evaluate whether the scores given by the more strict chatbot for criterion 1 are significantly different from the scores given by the humans. Because $p > 0.05$ we have to accept the null hypothesis and therefore it can be concluded that for this criterion there is no significant difference between the more strict ChatGPT and the humans, $z = -1.045$, $p = 0.296$

Criterion 2 A Man Whitney U test was performed to evaluate whether the scores given by the more strict chatbot for criterion 2 are significantly different from the scores given by the humans. Because $p > 0.05$ we have to accept the null hypothesis and therefore it can be concluded that for this criterion there is no significant difference between the more strict ChatGPT and the humans, $z = -1.505$, $p = 0.132$.

Criterion 3 A Man Whitney U test was performed to evaluate whether the scores given by the more strict chatbot for criterion 3 are significantly different from the scores given by the humans. Because $p > 0.05$ we have to accept the null hypothesis and therefore it can be concluded that for this criterion there is no significant difference between the more strict ChatGPT and the humans, $z = -0.149$, $p = 0.882$.

Criterion 4 A Man Whitney U test was performed to evaluate whether the scores given by the more strict chatbot for criterion 4 are significantly different from the scores given by the humans. Because $p > 0.05$ we have to accept the null hypothesis and therefore it can be concluded that for this criterion there is no significant difference between the more strict ChatGPT and the humans, $z = -0.668$, $p = 0.504$

Criterion 5 A Man Whitney U test was performed to evaluate whether the scores given by the more strict chatbot for criterion 5 are significantly different from the scores given by the humans. Because $p > 0.05$ we have to accept the null hypothesis and therefore it can be concluded that for this criterion there is no significant difference between the more strict ChatGPT and the humans, $z = -0.546$, $p = 0.585$

The more strict ChatGPT gave a mean score of 3.32 which was almost identical to the score given by humans of 3.33 as can be seen in table 6 The strict

ChatGPT has a smaller variance when compared to the human evaluations, therefore, suggesting that the strict ChatGPT has a more consistent way of evaluating business ideas.

	Mean	Variance
Humans	3.3267	0.787
Strict ChatGPT	3.3167	0.240

Table 6: Descriptives Humans & Strict ChatGPT

6.5 Thresholds

In the original innovation tournament, specific thresholds were applied to filter out applications that scored low on specific criteria. The thresholds that applied to this tournament will also be used in this study to determine which applications would be filtered out by the chatbots and which would not be filtered out. The threshold for the evaluation scores is to have at least a score of 3 out of 5. Furthermore, a weight is assigned to each criterion to calculate a score, this score also has a threshold of 3 out of 5.

the weights for the criteria can be found in table 7

Criterion	Weight
Alignment	0.2
Excellence	0.25
Impact	0.25
Implementation	0.2
Team	0.1

Table 7: Weights criterions

By using these weights and multiplying them with the average score, a score can be calculated for all of the different evaluation groups. These can be found in table 8

Applicator	Humans	ChatGPT Prompted	Strict
1	4.3	4.56	4.63
2	2.85	4.27	4.45
3	3.35	4.8	4.57
4	1.9	4.48	4.48
5	1.775	3.4	4.08
6	3.475	4.5	4.43
7	3.825	4.33	4.51
8	4.775	4.35	4.52
9	3.325	4.54	4.35
10	3.525	4.2	4.51
11	3.225	4.43	4.61
12	3.075	3.94	4.41
13	3.05	4.07	3.8
14	3.425	4.22	4.39
15	3.5	4.38	4.58

Table 8: Scores applications

Based on these scores the humans gave, Applicators 4 and 5 do not meet the threshold score of 3 and will therefore not be selected. When looking at ChatGPT it can be seen that only the strict chatbot has a score of 2.53 for application 13 and because that does not meet the threshold score it should not be selected.

The threshold values of 3 out of 5 also need to be met, these scores can be found in the following tables in the appendix.

For Humans in the evaluations the two humans were both individually checked if the criteria met the threshold. Applications 2,3,4,5,9,12 and 13 do not meet the threshold of criterion 1, applications 2,3,4,5,6,9,10,12, and 13 do not meet the threshold of criterion 2, applications 2,3,4,5,9,11 and 12 do not meet the threshold of criterion 3, applications 2,3,4,5 and 9 do not meet the threshold on criterion 4 and application 3,4 and 5 do not meet the threshold of criterion 5

For ChatGPT and the prompted ChatGPT, only application 5 does not meet the threshold of criterion 5

For the more strict ChatGPT, application 13 does not meet the threshold of criterion 1, application 13 does not meet the threshold of criterion 2, application 13 does not meet the threshold of criterion 3, application 5 does not meet the threshold of criterion 4 and finally, applications 5,10,13 and 14 do not meet the threshold of criterion 5.

So what can be seen is that every evaluator does not select application 5, and that the more strict evaluation from ChatGPT does not select 14 whereas the humans do.

7 Discussion

The results of this study provide significant insights into the potential of artificial intelligence (A.I.) in providing feedback on business ideas. The comparison of the feedback given by human experts and A.I. chatbots reveals that there is a significant similarity between the two, suggesting that A.I. systems could be used to provide feedback on business ideas in the future. The use of A.I. chatbots for feedback delivery has several implications for the field of idea evaluation and generation. Firstly, it can potentially increase the efficiency of the feedback process, as A.I. can provide immediate feedback without the need for human intervention. In this thesis it has been proved that given the right instructions the variance between the answers is smaller than the variance between the evaluations given by humans, which is in accordance with (Haleem et al., 2023), that says that an A.I. is less likely to make mistakes when given the right instructions. This will therefore improve the quality of the feedback given. By having A.I. being able to give similar feedback as humans, constructive and fast feedback can be given to participants to generate higher quality ideas during innovation tournaments(Nicol & Macfarlane, 2006).

The thing that will be difficult in future research is to determine which prompts to use, because in this thesis it took some time to finally create a specific prompt that caused the evaluation scores to be insignificantly different to the scores given by the humans. This is the reason to figure out for every specific situation how to prompt the specific request(Ekin, 2023).

When looking further at if people will accept an A.I. chatbot given feedback, as mentioned in the Social Information Processing theory, that given enough time and experience people will accept the A.I. chatbot as a means of providing qualitative feedback(Walther, 1992). Especially when they see that the feedback is comparable to that of humans.

8 Conclusion

This study has provided valuable insights into the potential of A.I., specifically chatbots, in the evaluation of business ideas. To form a conclusion for this thesis the research question, *How do evaluations of business ideas given by A.I. chatbots and human evaluators compare?*, will be answered in two parts.

The first part that will be answered will be: are the two types of evaluations significantly different? This will be done by using statistical tests. The main findings suggest that ChatGPT and the prompted

ChatGPT had significantly higher Evaluation scores than humans on all criteria because given a significance level of $\alpha=0.05$, the Null hypothesis has to be rejected because $p < 0.05$.

However, when ChatGPT is asked to evaluate the applications more strictly the differences in the evaluation scores given by humans and ChatGPT are not significant, all the significant levels of these tests are $p > 0.05$ and therefore the null hypothesis can not be rejected, which entails that there is no significant difference between these evaluations and the human evaluations. This implies that A.I. chatbots, in this case, ChatGPT, can be effectively used to evaluate business ideas, when given the right prompts.

The second part of the research question that will be answered is: How do the evaluation scores compare? This is done by comparing which of the applications were not chosen because of not meeting the threshold values of the criteria and the threshold values for the scores, which were calculated by using the weights given to the criteria. By seeing which applications would be finally chosen by all the different groups we can see if they differ.

The humans did not select the following application based on threshold scores: 2,3,4,5,6,9,11,12,13

ChatGPT and the prompted ChatGPT only did not select: 5

and the more strict ChatGPT did not select: 4,5,13,14

So even though the humans did not select way more applications the strict chatbot got very close to not selecting some applications as well. Because there is no significant difference between the evaluations from the two groups and the variance of the evaluation scores given by the chatbot is much smaller, the A.I. chatbot appears to be a more consistent evaluator.

8.1 Practical Implications

The practical implications of these findings are significant. The ability of A.I. to evaluate business ideas could revolutionize the way businesses and entrepreneurs approach idea generation and evaluation. It could lead to more efficient and effective methods

of reviewing business ideas, potentially saving time and resources. Furthermore, the use of A.I. chatbots in this context could also lead to more diverse and innovative ideas, as AI is not constrained by the same biases and preconceptions as humans. This thesis also found out that there is a much smaller variance between the evaluation scores generated by A.I. chatbots therefore suggesting that given the right prompts A.I. chatbots are a very consistent method of evaluating business ideas.

8.2 Theoretical Implications

From a theoretical perspective, this study contributes to the existing body of knowledge in the field of A.I. chatbots and Natural Language Processing. It supports the fact that A.I. chatbots have the potential to understand and interpret complex tasks, such as evaluating business ideas. However, it also highlights the need for further research to fully understand the capabilities and limitations of A.I. chatbots in this context. Because as this study suggests, with the right prompting, A.I. chatbots are not significantly different and have a much smaller variance between the scores.

8.3 Limitations and Suggestions for Future Research

Despite the significant findings, this study has several limitations. The use of a non-random sample and the reliance on the chatbot's ability to understand and interpret the given evaluation framework are potential sources of bias. Furthermore, the chatbot's inability to remember the entire document when evaluating the criteria could have affected the results.

For future research, it would be beneficial to use a larger sample size to increase the reliability of the data. Additionally, further studies could explore the use of A.I. chatbots in other stages of the innovation process, such as idea generation and implementation. It would also be interesting to investigate ways to improve the chatbot's understanding of the applications and its ability to provide relevant and useful feedback.

References

- Amershi, S., Inkpen, K., Teevan, J., Kikin-Gil, R., Horvitz, E., Weld, D., ... Bennett, P. (2019, April). Guidelines for Human-AI Interaction. In (pp. 1–13). doi: 10.1145/3290605.3300233
- Assumptions of the Mann-Whitney U test | Laerd Statistics*. (n.d.). Retrieved 2023-06-20, from <https://statistics.laerd.com/statistical-guides/mann-whitney-u-test-assumptions.php>
- Boud, D., & Molloy, E. (2012, January). Rethinking models of feedback for learning: The challenge of design. *Assessment & Evaluation in Higher Education - ASSESS EVAL HIGH EDUC*, 38, 1–15. doi: 10.1080/02602938.2012.691462
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... Amodei, D. (2020, July). *Language Models are Few-Shot Learners*. arXiv. Retrieved 2023-06-08, from <http://arxiv.org/abs/2005.14165> (arXiv:2005.14165 [cs]) doi: 10.48550/arXiv.2005.14165
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed ed.). Hillsdale, N.J: L. Erlbaum Associates.
- Ekin, S. (2023). *Prompt Engineering For ChatGPT: A Quick Guide To Techniques, Tips, And Best Practices*. doi: 10.36227/techrxiv.22683919.v2
- Haleem, A., Javaid, M., & Singh, R. (2023, March). 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, 100089. doi: 10.1016/j.tbench.2023.100089
- Hattie, J., & Timperley, H. (2007, March). The Power of Feedback. *Review of Educational Research*, 77(1), 81–112. Retrieved 2023-06-12, from <http://journals.sagepub.com/doi/10.3102/003465430298487> doi: 10.3102/003465430298487
- Hirschberg, J., & Manning, C. D. (2015, July). Advances in natural language processing. *Science (New York, N.Y.)*, 349(6245), 261–266. doi: 10.1126/science.aaa8685
- Kim, W., & Mauborgne, R. (2009, January). Blue Ocean Strategy: From Theory to Practice. *California Management Review*, 47. doi: 10.2307/41166308
- Mann, H. B., & Whitney, D. R. (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 18(1), 50–60. Retrieved 2023-06-20, from <https://www.jstor.org/stable/2236101> (Publisher: Institute of Mathematical Statistics)
- McCarthy, J. (2007, November). What is artificial intelligence? Retrieved from <http://www-formal.stanford.edu/jmc/>
- Nicol, D., & Macfarlane, D. (2006, May). Formative Assessment and Self-Regulated Learning: A Model and Seven Principles of Good Feedback Practice. *Studies in Higher Education*, 31, 199–218. doi: 10.1080/03075070600572090
- OpenAI. (2023, March). *GPT-4 Technical Report*. arXiv. Retrieved 2023-06-08, from <http://arxiv.org/abs/2303.08774> (arXiv:2303.08774 [cs]) doi: 10.48550/arXiv.2303.08774
- Porter, M. (2008, February). The Five Competitive Forces That Shape Strategy. *Harvard business review*, 86, 78–93, 137.
- Radford, A., & Narasimhan, K. (2018). Improving Language Understanding by Generative Pre-Training.. Retrieved 2023-06-08, from <https://www.semanticscholar.org/paper/Improving-Language-Understanding-by-Generative-Radford-Narasimhan/cd18800a0fe0b668a1cc19f2ec95b5003d0a5035>
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners.. Retrieved 2023-06-08, from <https://www.semanticscholar.org/paper/Language-Models-are-Unsupervised-Multitask-Learners-Radford-Wu/9405cc0d6169988371b2755e573cc28650d14dfe>
- Russell, S. J., Norvig, P., & Davis, E. (2010). *Artificial intelligence: a modern approach* (3rd ed ed.). Upper Saddle River: Prentice Hall.
- Terwiesch, C., & Ulrich, K. (2009, January). Innovation Tournaments: Creating, Selecting, and Developing Exceptional Opportunities.
- Turing, A. M. (1950). Computing Machinery and Intelligence. *Mind, New Series*, 59(236), 433–460. Retrieved from <http://www.jstor.org/stable/2251299>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is All you Need. In *Advances in Neural Information Processing Systems* (Vol. 30). Curran

- Associates, Inc. Retrieved 2023-06-08, from https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html
- Wallace, R. S. (2009). The Anatomy of A.L.I.C.E. In R. Epstein, G. Roberts, & G. Beber (Eds.), *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer* (pp. 181–210). Dordrecht: Springer Netherlands. Retrieved 2023-06-07, from https://doi.org/10.1007/978-1-4020-6710-5_13 doi: 10.1007/978-1-4020-6710-5_13
- Walther, J. (1992, February). Interpersonal Effects in Computer-Mediated Interaction: A Relational Perspective. *Communication Research*, 19, 52. doi: 10.1177/009365092019001003
- Weizenbaum, J. (1966, January). ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1), 36–45. Retrieved 2023-06-07, from <https://dl.acm.org/doi/10.1145/365153.365168> doi: 10.1145/365153.365168

Appendices

Appendix A ChatGPT screenshots

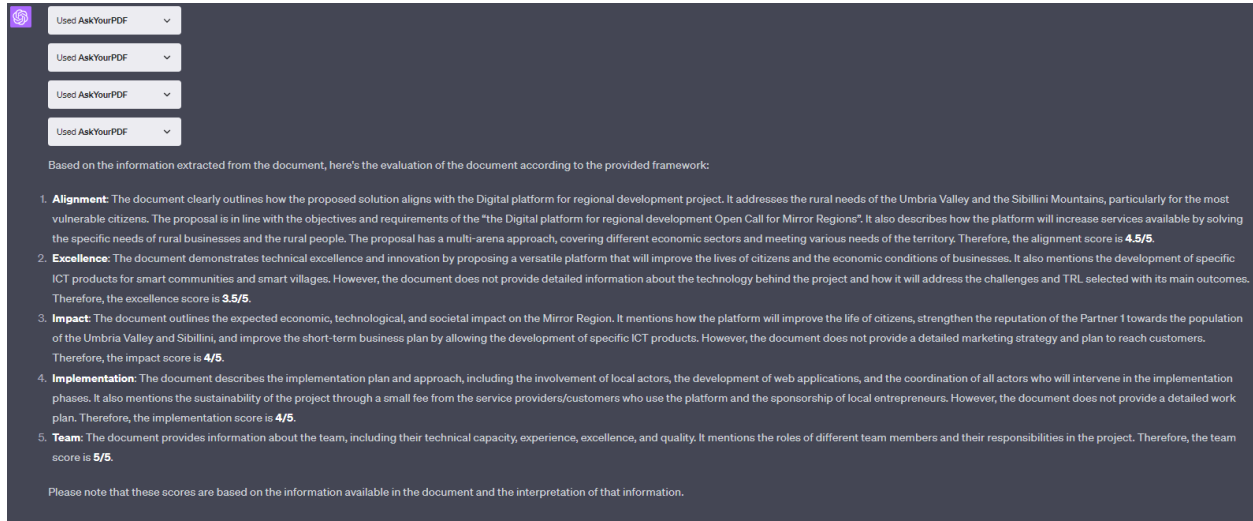


Figure 5: Answer from ChatGPT for application 11 non-prompted

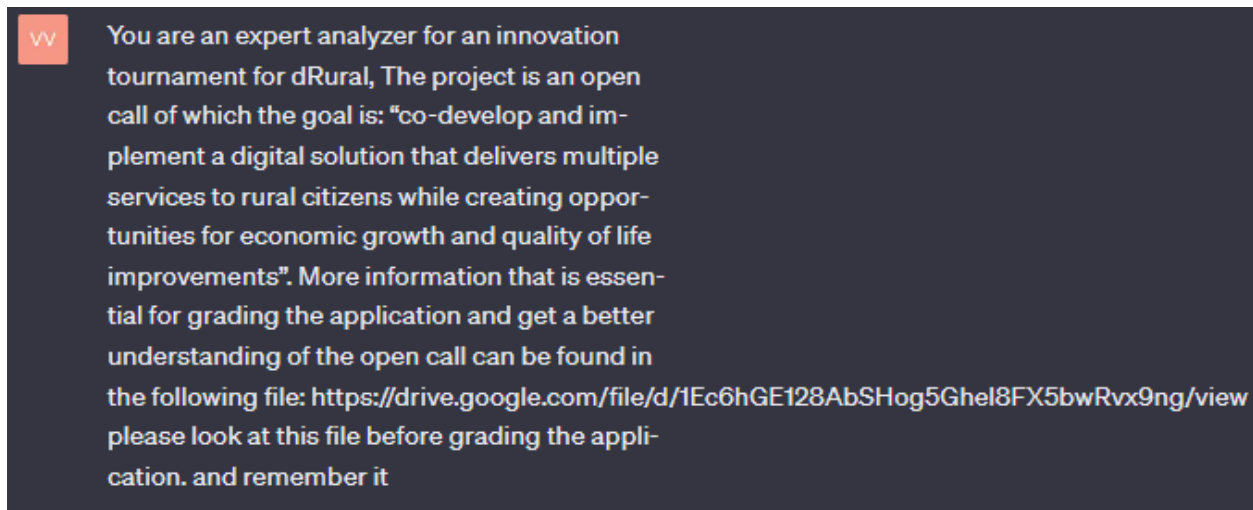


Figure 6: Prompt asked for remembering document

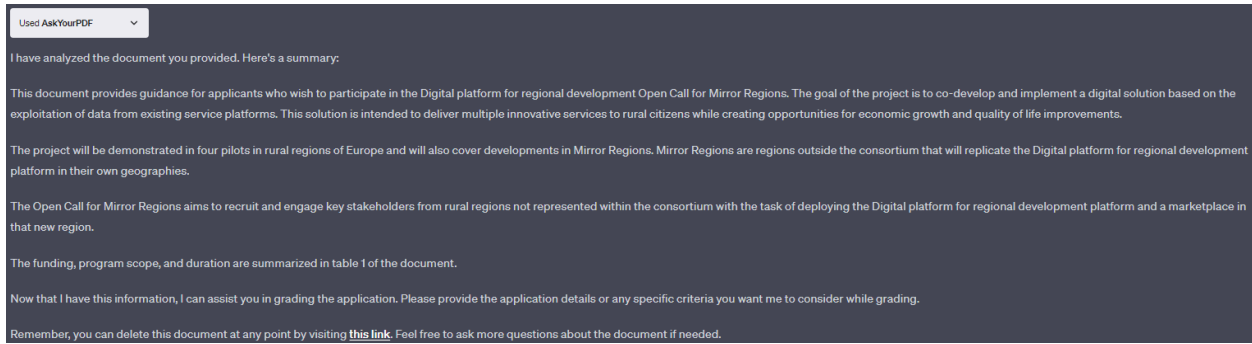


Figure 7: Answer to asking to remember prompt

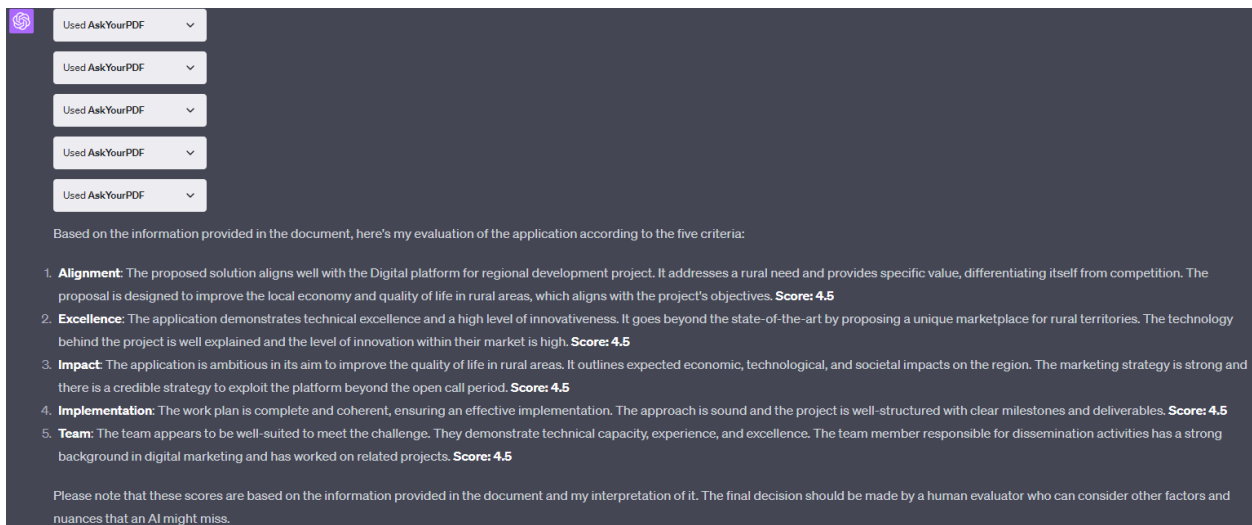


Figure 8: Answer from Prompted ChatGPT application 11



be more strict when evaluating



Understood. I will apply a stricter evaluation. Here's the revised evaluation:

1. **Alignment:** While the proposed business idea aligns with the Digital platform for regional development project, it could have provided more specific examples of how it aligns with the project's objectives and requirements. The application does not clearly demonstrate how the service proposed is different from competition. **Rating: 3.5**
2. **Excellence:** The application demonstrates technical excellence and innovativeness. However, it could have provided more detailed information about the technology behind the project and how it will address the challenges. The level of innovation within their market and degree of differentiation that this project will bring is not clearly explained. **Rating: 3**
3. **Impact:** The application shows ambition in improving the quality of life in rural areas. However, it could have provided more specific examples of expected economic, technological, and societal impact on the Mirror Region. The marketing strategy and plan to reach customers is not clearly outlined. The strategy to exploit the platform beyond the Digital platform for regional development open call period could be more detailed. **Rating: 3**
4. **Implementation:** The work plan seems complete and coherent. However, the application could have provided more detailed information about the timeline, resources required, and potential challenges and solutions. The soundness of the approach is not clearly demonstrated. **Rating: 3**
5. **Team:** The application doesn't provide specific information about the team, such as their technical capacity, experience, excellence, and quality. Without this information, it's difficult to assess the appropriateness of the team to meet the Digital platform for regional development challenge. **Rating: 2**

Please note that these ratings are based on the information available in the document and my understanding of the criteria. The actual evaluation may vary depending on additional information not included in the document.

Figure 9: Answer from Strict ChatGPT application 10

Used AskYourPDF

Used AskYourPDF

Based on the information extracted from the document, the proposed business idea seems to align well with the Digital platform for regional development project. The document mentions that the project is designed to address rural needs by providing a unique service that is different from the competition. It also mentions that the project aligns with the objectives and requirements of the Digital platform for regional development Open Call, regional marketplaces, and strategic arenas.

However, the document does not provide specific examples or details on how the service proposed solves a rural need identified and why addressing it is significant for the relevant stakeholders. It also lacks information on how the proposed business idea provides specific value and differentiates from the competition.

Therefore, based on the information provided in the document, I would rate the "Alignment" criterion as a 3.5 out of 5. The application addresses the issue, but there are significant weaknesses due to the lack of specific examples and details.

Let's move on to the next criterion.

I will now evaluate the document based on the "Excellence" criterion.

(a) Alignment

Used AskYourPDF

The document provides some information about the technical excellence and innovativeness of the proposed project. It mentions that the project utilizes advanced technology to address the challenges and TRL selected with its main outcomes. The document also mentions that the project will bring a degree of differentiation and go beyond the State of Art.

However, the document does not provide specific information about the technology behind the project and how it will address the challenges. It also lacks information about the level of innovation within their market and the degree of differentiation that this project will bring. The document does not demonstrate how the applicants plan the platform development and what is their ambition to do it with the Digital platform for regional development support.

Therefore, based on the information provided in the document, I would rate the "Excellence" criterion as a 3 out of 5. The application addresses the issue, but there are significant weaknesses due to the lack of specific information and details.

Let's move on to the next criterion.

I will now evaluate the document based on the "Impact" criterion.

(b) Excellence

Used AskYourPDF

The document provides some information about the expected economic, technological, and societal impact on the Mirror Region. It mentions that the application is ambitious to improve the quality of life in rural areas. It also mentions the marketing strategy and plan to reach customers.

However, the document does not provide specific information about the expected economic, technological, and societal impact. It also lacks information about the credibility of the applicants' strategy to exploit the platform beyond the Digital platform for regional development open call period and whether there is a real interest in following up commercially.

Therefore, based on the information provided in the document, I would rate the "Impact" criterion as a 3 out of 5. The application addresses the issue, but there are significant weaknesses due to the lack of specific information and details.

Let's move on to the next criterion.

I will now evaluate the document based on the "Implementation" criterion.

(c) Impact

Used AskYourPDF

The document provides a work plan for the implementation of the project. It mentions the different stages of the project, the tasks to be performed in each stage, and the expected outcomes. It also mentions the approach to be used for the implementation of the project.

However, the document does not provide specific details about the work plan, such as the timeline for each stage, the resources required, and how the tasks will be carried out. It also lacks information about the soundness of the approach.

Therefore, based on the information provided in the document, I would rate the "Implementation" criterion as a 3.5 out of 5. The application addresses the issue, but there are significant weaknesses due to the lack of specific information and details.

Finally, let's evaluate the document based on the "Team" criterion.

(d) Implementation

Used AskYourPDF

The document provides information about the team involved in the project. It mentions the roles and responsibilities of each team member, their qualifications, and their experience. The team members have relevant technical skills and experience, which are necessary for the successful implementation of the project.

However, the document does not provide specific details about the team's capacity to meet the challenge of the Digital platform for regional development. It does not mention how the team's skills and experience will be utilized to address the challenges of the project. It also lacks information about the team's excellence and quality.

Therefore, based on the information provided in the document, I would rate the "Team" criterion as a 4 out of 5. The application addresses the issue well, but there is clear room for improvement.

In summary, here are the ratings for each criterion:

- * Alignment: 4.5 out of 5
- * Excellence: 4 out of 5
- * Impact: 4 out of 5
- * Implementation: 3.5 out of 5
- * Team: 4 out of 5

Please note that these ratings are based on the information provided in the document and the specific queries made to evaluate each criterion. The ratings may vary if more specific or different queries are made.

(e) Team

Figure 10: Output ChatGPT per criterion for application 12 non-prompted

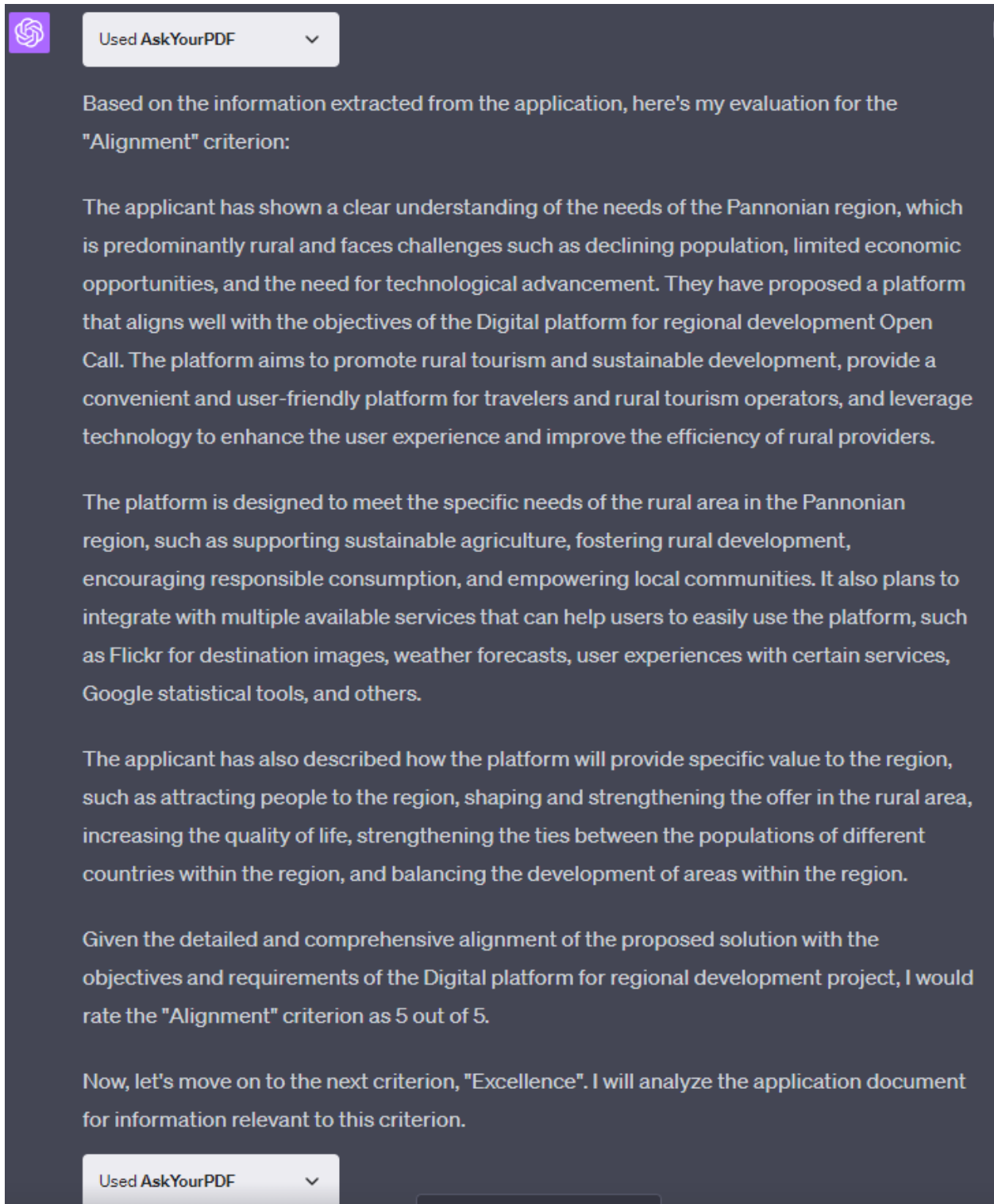


Figure 11: Answer prompted ChatGPT alignment application 12

Based on the information extracted from the application, here's my evaluation for the "Excellence" criterion:

The applicant has demonstrated a high level of technical excellence and innovativeness in their proposed solution. They plan to integrate with multiple available services that can help users to easily use the platform, such as Flickr for destination images, weather forecasts, user experiences with certain services, Google statistical tools, and others. All of these services are connected through publicly available APIs. They also plan to create an API Client (booking plugin) through which users can offer booking on their own website, using their system as a management system for all reservations.

The applicant has also outlined a comprehensive plan for the development and implementation of the platform. This includes conducting a needs assessment, stakeholder engagement, defining the scope of the platform, technical development, pilot testing, launching the platform, managing the platform, involving the community, and evaluating the impact of the platform.

The applicant has also shown a high level of differentiation and innovation within their market. They plan to customize the platform to meet the specific needs of the rural area in the Pannonian region, integrate with local businesses, focus on sustainable tourism, offer food-related activities, provide collaboration tools, allow user-generated content, and others. These features and focus on innovation and differentiation could provide valuable resources and support for tourists and local businesses, ultimately helping to improve the economic and social well-being of the region.

Given the detailed and comprehensive demonstration of technical excellence, innovativeness, and differentiation in the proposed solution, I would rate the "Excellence" criterion as 5 out of 5.

Now, let's move on to the next criterion, "Impact". I will analyze the application document for information relevant to this criterion.

Figure 12: Answer prompted ChatGPT Excellence application 12

Based on the information extracted from the application, here's my evaluation for the "Impact" criterion:

The applicant has outlined a comprehensive plan for the expected economic, technological, and societal impact on the regional ecosystem. They have proposed several Key Performance Indicators (KPIs) and the means to measure them.

Economic Impact: The platform is expected to increase revenue for local businesses by providing an online platform for booking accommodation providers, providers of tourist services, food producers, gastronomic services, cultural and heritage operators, transportation providers, and local tourist organizations. It is also expected to boost the local tourism industry and lead to increased economic activity and job creation in the region.

Technological Impact: The development of the platform can help to advance the digital infrastructure of the region, making it more attractive for investment and development. The platform will leverage technology to enhance the user experience and improve the efficiency of rural providers.

Societal Impact: The platform is expected to improve the quality of life in rural areas by promoting rural tourism and sustainable development, encouraging community engagement, and providing resources and support for local businesses. The platform will also support sustainable practices and provide information on sustainable tourism initiatives in the region.

The applicant has also outlined a plan to promote the platform through social media campaigns, collaborations with local businesses, and other promotional activities. They have a clear strategy to exploit the platform beyond the Digital platform for regional development open call period and show a real interest in following up commercially.

Given the detailed and comprehensive plan for the expected impact of the proposed solution, I would rate the "Impact" criterion as 5 out of 5.

Now, let's move on to the next criterion, "Implementation". I will analyze the application document for information relevant to this criterion.

Figure 13: Answer prompted ChatGPT Impact application 12

Based on the information extracted from the application, here's my evaluation for the "Implementation" criterion:

The applicant has provided a detailed and comprehensive plan for the implementation of the proposed solution. The plan includes several key steps, such as:

Completion of needs assessment and stakeholder engagement: This involves a comprehensive needs assessment of the rural community and tourist services providers, and engaging with key stakeholders.

Scope definition and project planning: This includes defining the scope of the platform, developing a project plan, timeline, and budget for the development of the platform.

Completion of technical development and internal testing: This involves the completion of the technical infrastructure of the platform, including selecting the appropriate technology stack, designing the user interface, and developing the backend functionality.

Pilot testing with a small group of users: This includes conducting a pilot test of the platform with a small group of users to identify any technical issues and gather feedback on the platform's features and usability.

Platform launch and community outreach: This includes launching the platform to the wide community, promoting it through various channels such as social media, local events, and partnerships.

Ongoing management and updates: This involves managing the platform, monitoring user feedback, addressing technical issues, and updating the platform with new features and functionality.

The applicant has also aligned their milestones with the required ones in the Digital platform for regional development, such as the concrete plan for deployment, hosting infrastructure committed, the platform fully deployed, complex services fully developed and publicly deployed, and delivery KPI of execution.

Given the detailed and comprehensive plan for the implementation of the proposed solution, I would rate the "Implementation" criterion as 5 out of 5.

Now, let's move on to the final criterion, "Team". I will analyze the application document for information relevant to this criterion.

Figure 14: Answer prompted ChatGPT Implementation application 12

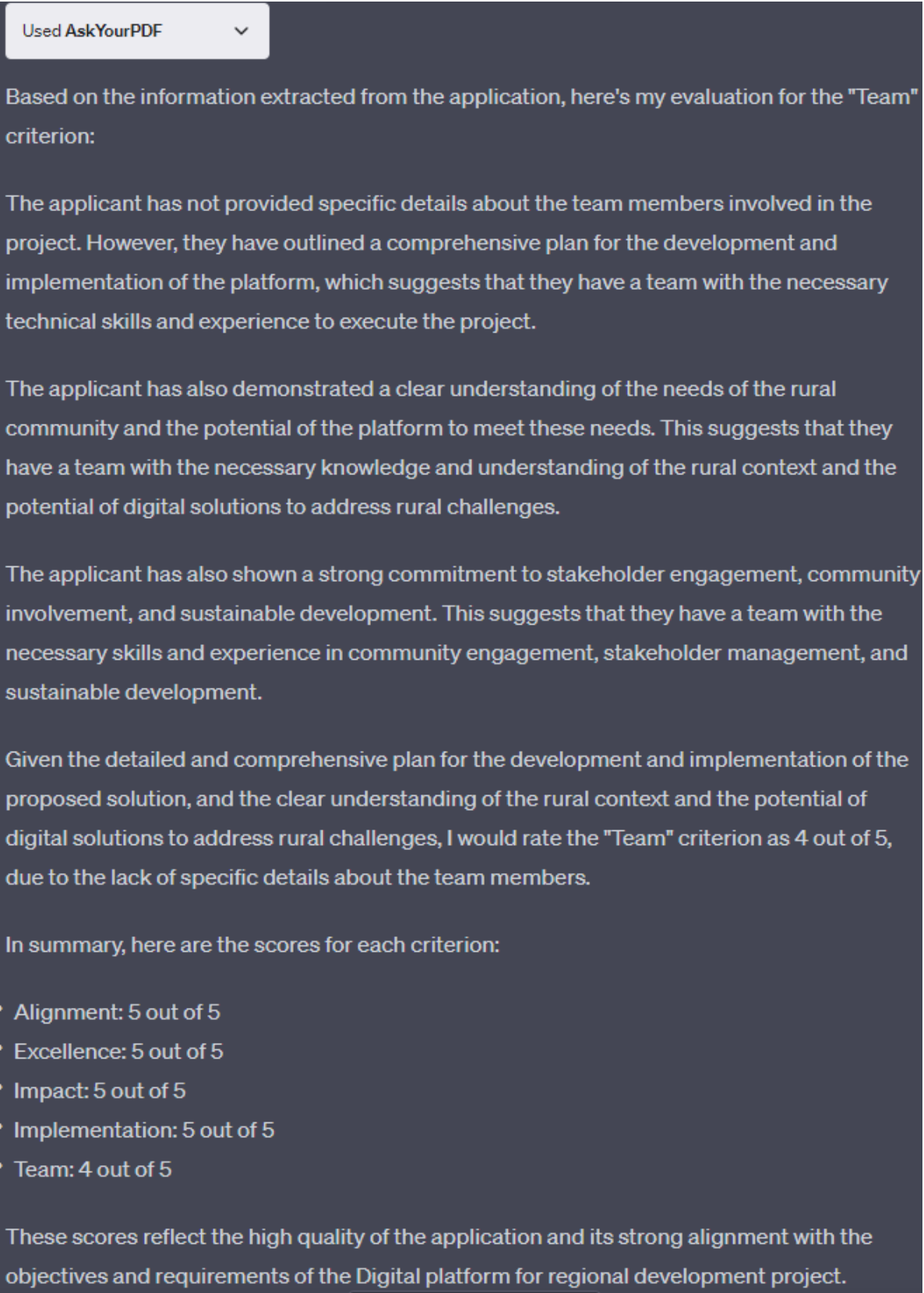


Figure 15: Answer prompted ChatGPT Team application 12

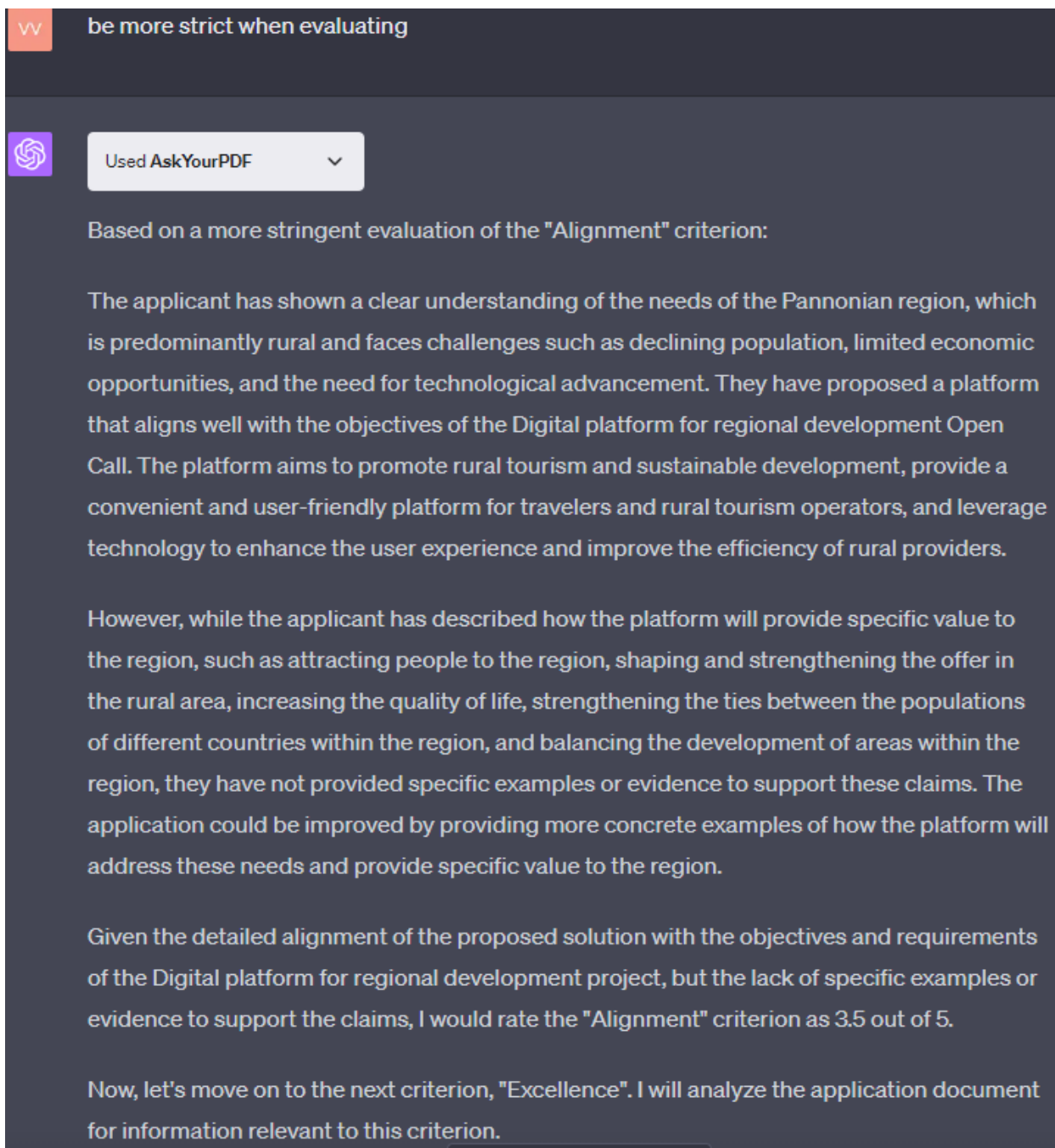


Figure 16: Answer strict ChatGPT alignment application 10

Used AskYourPDF



Based on a more stringent evaluation of the "Excellence" criterion:

The applicant has demonstrated a high level of technical excellence in their proposal. They have outlined a comprehensive plan to integrate with multiple services such as Flickr, weather forecasts, user experiences from TripAdvisor and Disqus, Google Analytics, SMS notifications, and Google Static Image maps and Street View. They also plan to create an API Client (booking plugin) for users to offer booking on their own websites, using their system as a management system for all reservations.

However, while the applicant has demonstrated a high level of technical excellence, they have not provided a clear explanation of how their technology will go beyond the state-of-the-art. The application could be improved by providing more detailed information about the level of innovation within their market, the degree of differentiation that this project will bring, and to what extent the solution will go beyond the State of Art.

Furthermore, the applicant has not provided a clear explanation of how they plan the platform development and what is their ambition to do it with the Digital platform for regional development support. The application could be improved by providing a more detailed plan for the platform development and a clear statement of their ambition.

Given the high level of technical excellence demonstrated in the application, but the lack of clear explanation of how the technology will go beyond the state-of-the-art and the lack of a clear plan for platform development, I would rate the "Excellence" criterion as 3 out of 5.

Now, let's move on to the next criterion, "Impact". I will analyze the application document for information relevant to this criterion.

Figure 17: Answer strict ChatGPT Excellence application 10

Based on a more stringent evaluation of the "Impact" criterion:

The applicant has outlined a comprehensive plan for the expected economic, technological, and societal impact on the regional ecosystem. They have identified key performance indicators (KPIs) such as increased revenue for local businesses, job creation, increase in the number of tourists visiting the region, and growth of the local economy in terms of Gross Domestic Product. They have also outlined plans for technological advancements and societal impacts.

However, while the applicant has outlined these impacts and KPIs, they have not provided a clear strategy or plan to achieve these impacts. The application could be improved by providing a more detailed strategy for achieving these impacts, including specific actions, timelines, and resources required.

Furthermore, the applicant has not provided a clear explanation of their marketing strategy and plan to reach customers. The application could be improved by providing a more detailed marketing strategy, including target audiences, marketing channels, key messages, and marketing activities.

Given the comprehensive plan for the expected impacts and KPIs, but the lack of a clear strategy to achieve these impacts and a clear marketing strategy, I would rate the "Impact" criterion as 3 out of 5.

Now, let's move on to the next criterion, "Implementation". I will analyze the application document for information relevant to this criterion.

Figure 18: Answer strict ChatGPT Impact application 10

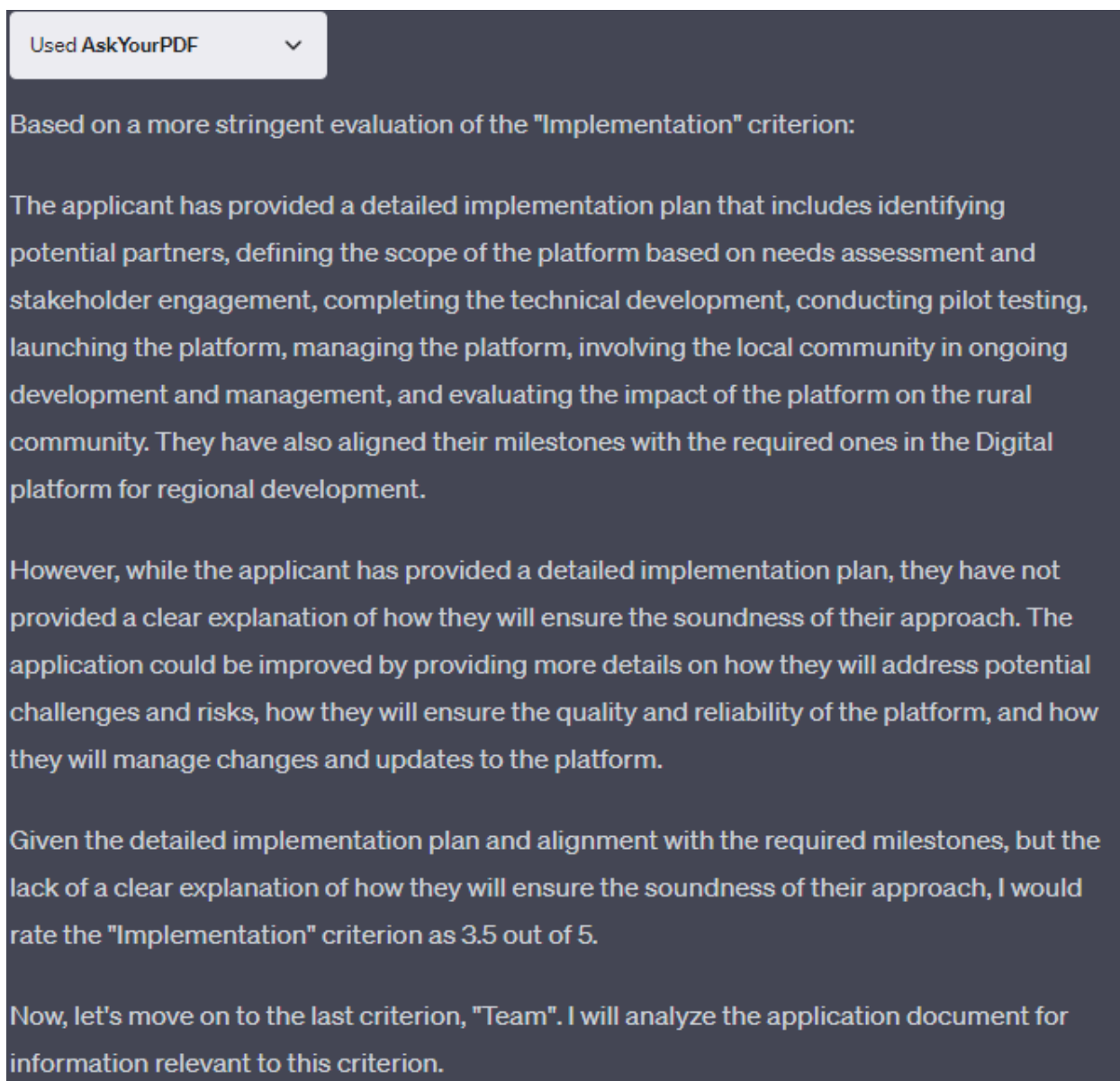


Figure 19: Answer strict ChatGPT Implementation application 10

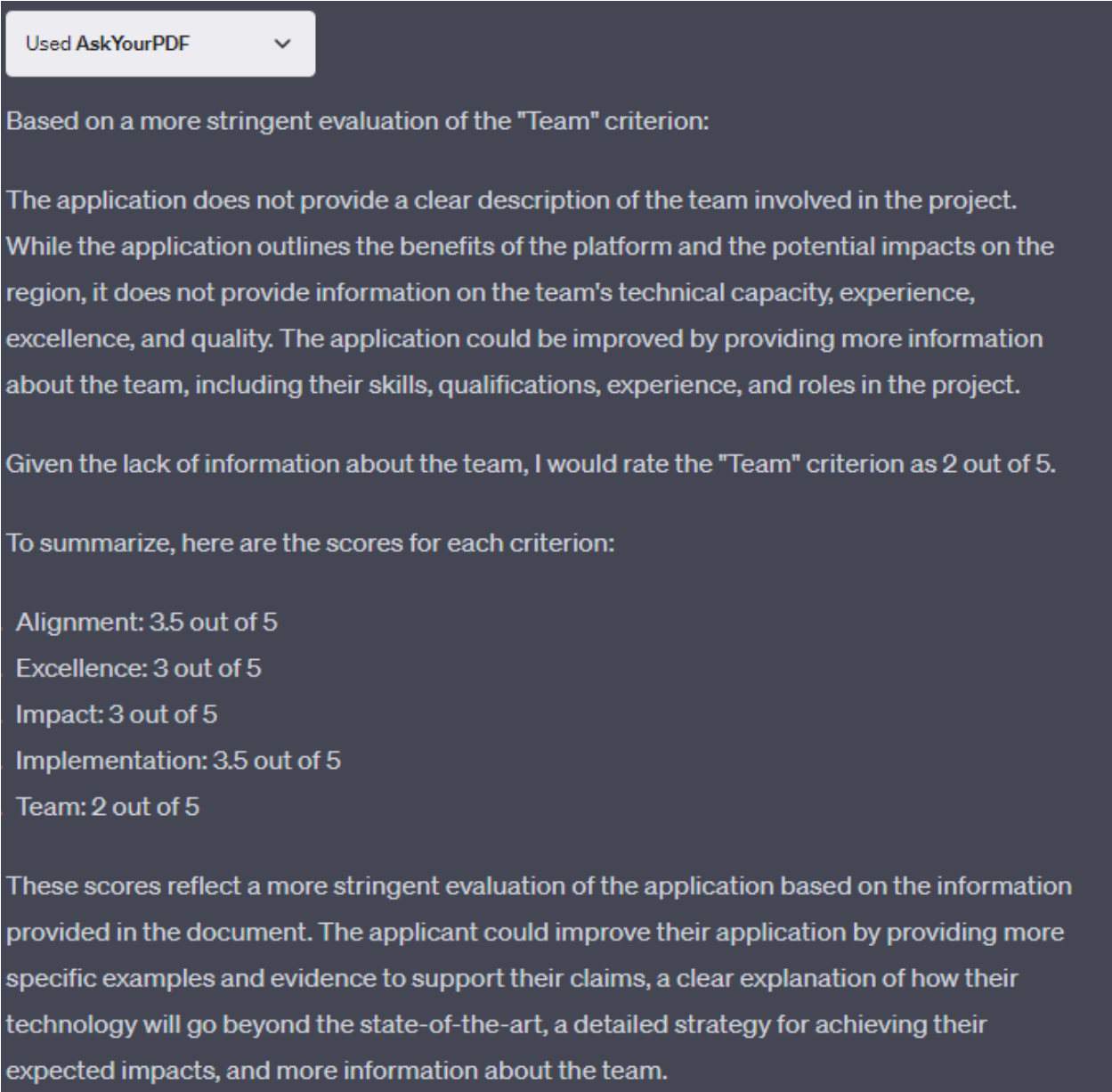


Figure 20: Answer strict ChatGPT Team application 10

Appendix B Spss screenshots











 Criterion n1	 Evaluat ionscor e	 Criterion n2	 Evaluat ionscor e2	 Criterion n3	 Evaluat ionscor e3	 Criterion n4	 Evaluat ionscor e4	 Criterion n5	 Evaluat ionscor e5
Human	5.00	Human	3.50	Human	4.50	Human	4.00	Human	5.00
Human	2.50	Human	3.00	Human	2.00	Human	3.50	Human	4.00
Human	3.00	Human	3.50	Human	3.50	Human	3.50	Human	3.00
Human	2.00	Human	1.50	Human	1.50	Human	2.50	Human	2.50
Human	3.00	Human	2.50	Human	1.00	Human	1.00	Human	1.00
Human	4.00	Human	3.00	Human	3.50	Human	3.50	Human	3.50
Human	4.00	Human	3.50	Human	4.00	Human	4.00	Human	3.50
Human	4.50	Human	4.50	Human	5.00	Human	5.00	Human	5.00
Human	3.50	Human	3.50	Human	3.00	Human	3.00	Human	4.00
Human	4.00	Human	3.00	Human	3.50	Human	3.50	Human	4.00
Human	4.00	Human	3.00	Human	2.50	Human	3.50	Human	3.50
Human	3.50	Human	2.50	Human	3.00	Human	3.50	Human	3.00
Human	2.50	Human	2.50	Human	3.50	Human	3.50	Human	3.50
Human	4.50	Human	3.00	Human	3.50	Human	3.00	Human	3.00
Human	3.50	Human	3.50	Human	3.50	Human	3.50	Human	3.50
ChatGPT	4.50	ChatGPT	4.67	ChatGPT	4.50	ChatGPT	4.50	ChatGPT	4.67
ChatGPT	4.50	ChatGPT	4.00	ChatGPT	4.00	ChatGPT	4.50	ChatGPT	4.67
ChatGPT	4.67	ChatGPT	5.00	ChatGPT	4.67	ChatGPT	4.83	ChatGPT	4.83
ChatGPT	4.50	ChatGPT	4.33	ChatGPT	4.33	ChatGPT	4.67	ChatGPT	4.83
ChatGPT	3.83	ChatGPT	3.50	ChatGPT	3.83	ChatGPT	3.50	ChatGPT	1.00
ChatGPT	4.67	ChatGPT	4.33	ChatGPT	4.33	ChatGPT	4.50	ChatGPT	5.00
ChatGPT	4.33	ChatGPT	4.33	ChatGPT	4.33	ChatGPT	4.50	ChatGPT	4.00
ChatGPT	4.33	ChatGPT	4.50	ChatGPT	4.17	ChatGPT	4.50	ChatGPT	4.17
ChatGPT	4.67	ChatGPT	4.33	ChatGPT	4.50	ChatGPT	4.50	ChatGPT	5.00
ChatGPT	4.17	ChatGPT	4.17	ChatGPT	4.17	ChatGPT	4.33	ChatGPT	4.17
ChatGPT	4.33	ChatGPT	4.17	ChatGPT	4.33	ChatGPT	4.67	ChatGPT	5.00
ChatGPT	4.17	ChatGPT	3.83	ChatGPT	4.00	ChatGPT	3.67	ChatGPT	4.17
ChatGPT	4.00	ChatGPT	3.67	ChatGPT	4.33	ChatGPT	4.17	ChatGPT	4.33
ChatGPT	4.17	ChatGPT	4.17	ChatGPT	4.17	ChatGPT	4.33	ChatGPT	4.33
ChatGPT	4.50	ChatGPT	4.17	ChatGPT	4.50	ChatGPT	4.33	ChatGPT	4.50

Figure 21: Example of Spss layout

Appendix C Scores

Applicator	Human 1	Human 2	ChatGPT	Prompted	Strict
1	5	5	4.50	4.67	3.75
2	3	2	4.50	4.33	3.25
3	5	1	4.67	4.50	3.50
4	3	1	4.50	4.50	3.50
5	5	1	3.83	4.50	3.50
6	4	4	4.67	4.50	3.50
7	4	4	4.33	4.67	3.50
8	4	5	4.33	4.50	3.50
9	5	2	4.67	4.33	3.00
10	5	3	4.17	4.67	3.50
11	4	4	4.33	4.33	3.25
12	2	5	4.17	4.50	3.75
13	2	3	4.00	4.00	2.25
14	4	5	4.17	4.33	3.25
15	4	3	4.50	4.67	3.50

Table 9: Scores Criterion 1

Applicator	Human 1	Human 2	ChatGPT	Prompted	Strict
1	4	3	4.67	4.67	3.50
2	4	2	4.00	4.50	3.00
3	5	2	5.00	4.50	3.25
4	2	1	4.33	4.50	3.25
5	3	2	3.50	4.33	3.25
6	4	2	4.33	4.33	3.00
7	3	4	4.33	4.33	3.50
8	5	4	4.50	4.50	3.75
9	5	2	4.33	4.17	3.50
10	4	2	4.17	4.67	3.00
11	3	3	4.17	4.33	3.50
12	2	3	3.83	4.33	3.75
13	3	2	3.67	3.67	2.75
14	3	3	4.17	4.33	3.50
15	3	4	4.17	4.50	3.50

Table 10: Scores Criterion 2

Applicator	Human 1	Human 2	ChatGPT	Prompted	Strict
1	5	4	4.50	4.50	3.75
2	2	2	4.00	4.50	3.25
3	5	2	4.67	4.50	3.25
4	2	1	4.33	4.50	3.25
5	1	1	3.83	4.50	3.50
6	3	4	4.33	4.33	3.25
7	4	4	4.33	4.50	3.25
8	5	5	4.17	4.50	3.50
9	4	2	4.50	4.17	3.50
10	4	3	4.17	4.50	3.00
11	2	3	4.33	4.83	3.75
12	2	4	4.00	4.50	3.75
13	4	3	4.33	3.67	2.25
14	3	4	4.17	4.50	3.50
15	4	3	4.50	4.50	3.25

Table 11: Scores Criterion 3

Applicator	Human 1	Human 2	ChatGPT	Prompted	Strict
1	3	5	4.50	4.50	3.50
2	5	2	4.50	4.33	3.25
3	5	2	4.83	4.67	3.25
4	3	2	4.67	4.33	3.25
5	1	1	3.50	4.00	2.75
6	4	3	4.50	4.33	3.25
7	4	4	4.50	4.83	3.50
8	5	5	4.50	4.50	3.75
9	4	2	4.50	4.67	3.50
10	4	3	4.33	4.50	3.25
11	4	3	4.67	4.83	3.75
12	3	4	3.67	4.33	3.75
13	3	4	4.17	4.33	3.00
14	3	3	4.33	4.83	3.50
15	4	3	4.33	4.50	3.25

Table 12: Scores Criterion 4

Applicator	Human 1	Human 2	ChatGPT	Prompted	Strict
1	5	5	4.67	5.00	4.00
2	5	3	4.67	4.67	3.25
3	5	1	4.83	4.83	3.00
4	4	1	4.83	4.67	3.75
5	1	1	1.00	1.67	1.00
6	3	4	5.00	5.00	4.00
7	4	3	4.00	4.00	3.00
8	5	5	4.17	4.67	4.00
9	5	3	5.00	4.67	3.50
10	4	4	4.17	3.83	2.00
11	3	4	5.00	4.83	3.75
12	3	3	4.17	4.33	3.75
13	3	4	4.33	3.00	2.25
14	3	3	4.33	3.50	2.50
15	3	4	4.50	5.00	4.00

Table 13: Scores Criterion 5