HCI evaluation based on AB testing: ChatGPT and Bard

RUI ZHU r.zhu.22@abdn.ac.uk

Abstract:

Assessing human-computer interaction (HCI) is crucial in the field of technological progress to enhance the collaboration between technological systems and user involvement (Ren, Silpasuwanchai & Cahill 2019). This essay aims to comprehensively assess two conversational artificial intelligence platforms, namely ChatGPT and Bard, using an unbiased method. The focus is on utilizing A/B testing is a method of using qualitative and quantitative data collection to compare and analyze the performance of tools in human-computer interaction (Lazar Feng & Hochheiser 2017). System Usability Scale (SUS) questionnaire scores are then used to understand usability, interface design, and the overall user experience. The paper extensively analyzes multiple aspects of user feedback and provides a detailed and accurate assessment based on scores collected from the SUS questionnaire and a thorough review of the study design, methodology, implementation techniques, and analysis of results. It provides crucial insights for future assessments of human-computer interaction. Finally, the data content is comprehensively explained and the results are obtained through the research methods, technology, implementation process and result analysis to facilitate the future evaluation of human-computer interaction.

Keywords

AB Testing, Conversational Artificial Intelligence, Performance Comparison, Human-Computer Interaction, Improvement Strategies, System Usability Scale (SUS)

Introduction

Due to the increasing prominence of artificial intelligence technology, conversational AI systems are becoming increasingly important in both everyday life and the workplace. The systems could imitate natural language interaction, resulting in enhanced communication between humans and computers (Karat and Karat 2003). The aim of this study is to evaluate the human-computer interaction of OpenAI's conversational AI tools, ChatGPT and Bard, using the A/B test approach together with qualitative user feedback and quantitative SUS questionnaire ratings. The review aims to assess the merits and drawbacks of both goods in relation to user experience, providing significant perspectives for enhancing future products. The report offers a thorough evaluation of both tools in relation to functionality, ease of use, efficiency, reliability, and user satisfaction. The results of the SUS survey show that ChatGPT is more usable and has a higher level of user satisfaction as compared to Bard. The research data method and the issue of user experience are some of the most critical aspects of the human-computer interaction, an area that deals specifically with how people interact and use computer systems. Analyzing the user experience will help us understand the strong points and weak points of the system thereby enhancing it so as to increase user satisfaction.

Application description

ChatGPT and Bard are two conversational artificial intelligence tools designed for human-computer interaction, designed to provide users with a convenient and natural interactive experience. ChatGPT focuses on the application of natural language processing technology to provide a conversation experience closer to human communication (Mattas 2023). Bard is characterized by its rich and diverse functional characteristics, and strives to provide users with a variety of application scenarios.

The evaluation focused on the user interface, interaction flows and overall user satisfaction of the two products. Such an evaluation is necessary to understand how the user interacts with the tool and what its benefits and further improvements can be achieved. Interface design involves the visual layout of tools, the arrangement of elements, and the ease of user operation (Tidwell 2010). Interaction process includes the issues of information exchange between the user and the tool, the response time of the system, and the operation comfort. Overall user satisfaction means how much user considers a tool useful depending upon his own preferences, responses of others, and overall ratings for an application. The examination of these major issues will provide a complete analysis on how effectively these two instruments relate with human–computer interaction.

Evaluation Strategy

This assessment is designed to answer specific questions about the comparative analysis between ChatGPT and Bard. The research aims at analysing and explaining the differences between these tools in terms of user experience, interface design and interaction flows with respect to certain characteristics that affect the overall user experience. A comparative analysis will be made on user experience aspects while understanding the differences in interface design, interaction flow and user satisfaction for both chats. These goals will be achieved through a structured approach: Goal 1 involves a detailed comparison of the interface design and interaction flow of the two tools, while Goal 2 requires an in-depth analysis of user satisfaction metrics to measure usability. Goal 3 will delve into the specific features that affect the ChatGPT and Bard user experience. The expected deliverables include a detailed report outlining a comparative analysis that provides insights into the strengths of each tool in different ways, which will pave the way for potential future enhancements. A variety of assessment methods are considered, such as heuristic assessment and user interviews; However, the A/B test method was chosen because it provides a controlled environment for direct and measurable comparison of the performance of the two tools. This approach provides a unique combination of quantitative and qualitative insights, enabling a comprehensive assessment of user interaction and experience.

Research methodology

We chose A/B testing as the primary method of evaluation, and each tool invited 10 users to participate. The selection of these particular participants contributed to ensuring authenticity towards evaluating experiences which are aimed at comparing ChatGPT and Bard comprehensively towards their efficacy and efficiency outcomes. Ten well-designed questions were used in this comprehensive assessment of participants' experiences with the two products. The following methods and standards were employed to ensure the validity and reliability of the research:

Selection of participants: A total of ten participants were involved in this A/B test. The participants were students from the University of Aberdeen, friends and family members, and provided a detailed review of Chat GTP and Bard.

Study size: In terms of study design, we considered involving a maximum of ten to thirty participants so that the information obtained could be representative and reliable.

Limiting the study scope: This evaluation will not involve public consultation. Instead, it will be guided by guidelines of research ethics and respect the privacy of participants.

Scoring mechanism: For each participant, all questions should be scored on a scale of 0 to 10, with 10

indicating complete satisfaction and 0 indicating extreme dissatisfaction.

How we will process the data: We will compare the average scores for Chat GPT and Bard for each question in order to identify and analyse their strengths and weaknesses.

Thus, this research will attempt to understand the detailed views of ChatGPT and Bard on various aspects of the user experience for this approach. We hope that these observations will help us to know the exact user needs of such tools and suggest their modifications. By conducting a thorough analysis, we hope to provide meaningful direction and stimulation for the future development of artificial intelligence-based conversational systems.

1. Interface Design: Is ChatGPT/Bard's interface designed to make it easy for you to find the features you need?
2. Smoothness of Operation: Does the interface flow smoothly and without delay when using ChatGPT/Bard?
3. Dialogue Interaction: Do the dialogue and answer functions in ChatGPT/Bard meet your needs?
4. Functionality: Is the functionality of Bard better suited to your needs than that of ChatGPT?
5. User Experience: Do the colours and layout design meet your expectations when using ChatGPT/Bard?
6. Performance efficiency: Does the performance of ChatGPT/Bard meet your expectations?
7. File Handling: Does ChatGPT/Bard meet your needs in terms of file saving, exporting and sharing?
8. Uniqueness: Does ChatGPT/Bard have unique features compared to other online dialogue tools you have used?
9. Usage Preference: Would you like to use ChatGPT/Bard more than other online dialogue tools?
10. Satisfaction: Were you able to create a satisfying interaction experience with ChatGPT/Bard in the time frame expected?

Through this process, participants provided qualitative feedback, offering insights into the tools' usability, interactivity, and specific functionalities, shedding light on their overall experiences.

In addition, to measure users' perceived usability of the tool more comprehensively, we specifically used the SUS (System Usability Scale) questionnaire. This quantitative survey tool provides users with a quantitative assessment of tool usability, including user perceptions of interface friendliness, operational complexity, and overall satisfaction. For instance, in text-based interactions, ChatGPT showcases a more succinct interface design, emphasizing direct question-answer interactions (Khennouche et al. 2023). Conversely, Bard offers a richer array of interactive elements and diverse response formats, occasionally presenting information visually, such as through tables. Regarding sharing functionalities, Bard surpasses ChatGPT by offering a broader spectrum of sharing options, including exporting to various file formats or creating drafts in emails, whereas ChatGPT primarily generates shareable links. Bard stands out for offering multiple alternative answers at times, diversifying user interaction. Contrastingly, ChatGPT tends to present a single response, maintaining a more streamlined interaction experience (Ray 2023). The Bard's interface, while rich in functions, can be intricate, potentially leaving users feeling that navigation isn't intuitive or operations are complex. In contrast, ChatGPT prioritizes simplicity, emphasizing direct text interactions for a more user-friendly experience. The focus of the evaluation is to explore the advantages and disadvantages of both tools in terms of user experience. These evaluations are designed to provide a comprehensive reference for enhancing product features and advancing development. Through this ongoing analysis, we try to gather deeper insights that will help improve and optimize ChatGPT and Bard. Our goal is to effectively meet user needs and ensure that both tools provide a seamless and engaging interactive experience.

Implementation process

In the implementation process, the A/B testing direct comparative strategy was used. Providing a detailed examination of many iterations of the tools such as ChatGPT and Bard. During the participants interacted with either ChatGPT or Bard and gave comprehensive feedback on them. This feedback captures their feelings, views, and ratings on the particular tool's functions observed while using them. Through the compilation of qualitative feedback from users, we have gained a comprehensive understanding of their perceptions and experiences with the product. These factors include the usability, capacity, fluency, and dialogic quality of the interface.

In the paired samples statistical analysis, we conducted an analysis between the performance of ChatGPT (A condition) and Bard (B condition). Figure 1 demonstrates that ChatGPT consistently obtained higher ratings in all ten sets of comparisons, as indicated by the mean user ratings for each scenario. The average ratings for the paired samples varied between 5.67 and 9.00 (condition A) and between 5.56 and 8.67 (condition B), with standard deviations ranging from 0.833 to 2.291, showing diversity in the ratings. The standard errors vary between 0.278 and 0.764, providing significant insights into the accuracy of the mean estimate.

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	A1	9.00	9	1.000	.333
	B1	7.33	9	1.414	.471
Pair 2	A2	8.11	9	.928	.309
	B2	5.56	9	2.007	.669
Pair 3	A3	8.44	9	1.130	.377
	B3	6.44	9	2.242	.747
Pair 4	A4	8.00	9	1.000	.333
	B4	6.44	9	1.590	.530
Pair 5	A5	8.56	9	1.014	.338
	B5	6.33	9	1.936	.645
Pair 6	A6	8.44	9	1.014	.338
	B6	6.67	9	1.871	.624
Pair 7	A7	7.78	9	.833	.278
	B7	6.56	9	2.007	.669
Pair 8	A8	8.67	9	1.225	.408
	B8	7.00	9	2.345	.782
Pair 9	A9	8.22	9	.972	.324
	B9	7.11	9	1.833	.611
Pair 10	A10	9.00	9	1.225	.408
	B10	5.67	9	2.291	.764

Figure 1

As shown in Figure 2, we further examined the correlation between ChatGPT and the Bard score. The correlation coefficients range from -0.505 to 0.531, showing a change from a moderate negative correlation to a moderate positive correlation. This suggests that in some cases the trends in user ratings for ChatGPT are opposite to those for Bard, while in other cases they converge.

Paired Samples Correlations

		N	Correlation	Significance One-Sided p	Significance Two-Sided p
Pair 1	A1 & B1	9	.177	.325	.649
Pair 2	A2 & B2	9	-.239	.268	.536
Pair 3	A3 & B3	9	.356	.173	.347
Pair 4	A4 & B4	9	-.157	.343	.686
Pair 5	A5 & B5	9	.403	.141	.282
Pair 6	A6 & B6	9	-.505	.083	.165
Pair 7	A7 & B7	9	.531	.070	.141
Pair 8	A8 & B8	9	.131	.369	.738
Pair 9	A9 & B9	9	-.156	.344	.689
Pair 10	A10 & B10	9	-.356	.173	.347

Figure 2

The results of the paired samples t-test, shown in Figure 3, provide further evidence of the significance of the difference between ChatGPT and Bard. We noticed that several pairs showed statistically significant score differences. In particular, the two-sided p-values for pairs 1,2, 3,4, 5, 6 and 10 are below

the commonly used significance level of 0.05, indicating that the performance of ChatGPT is significantly different from Bard in these pairs. The 95% confidence intervals do not cross zero, further confirming the existence of these differences.

Paired Samples Test

		Paired Differences					t	df	Significance	
					95% Confidence Interval of the Difference					
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper			One-Sided p	Two-Sided p
Pair 1	A1 - B1	1.667	1.581	.527	.451	2.882	3.162	8	.007	.013
Pair 2	A2 - B2	2.556	2.404	.801	.708	4.403	3.190	8	.006	.013
Pair 3	A3 - B3	2.000	2.121	.707	.369	3.631	2.828	8	.011	.022
Pair 4	A4 - B4	1.556	2.007	.669	.013	3.098	2.325	8	.024	.049
Pair 5	A5 - B5	2.222	1.787	.596	.848	3.596	3.730	8	.003	.006
Pair 6	A6 - B6	1.778	2.539	.846	-.174	3.729	2.101	8	.034	.069
Pair 7	A7 - B7	1.222	1.716	.572	-.097	2.541	2.137	8	.033	.065
Pair 8	A8 - B8	1.667	2.500	.833	-.255	3.588	2.000	8	.040	.081
Pair 9	A9 - B9	1.111	2.205	.735	-.584	2.806	1.512	8	.085	.169
Pair 10	A10 - B10	3.333	2.958	.986	1.060	5.607	3.381	8	.005	.010

Figure 3

Finally, Figure 4 shows the effect size analysis for each pairing, as measured by Cohen's d and Hedges' g. Cohen's d values range from 0.700 to 2.812, while Hedges' g, slightly adjusted to account for the small sample size, also shows similar trends. Large effect size metrics, such as those observed in pairs 2, 4, 6 and 8, indicate that the performance difference between ChatGPT and Bard is not only statistically significant, but also has important implications in practice. The 95% confidence intervals for these effect sizes provide a range of reliability for the difference estimates, adding confidence to our conclusions.

Paired Samples Effect Sizes

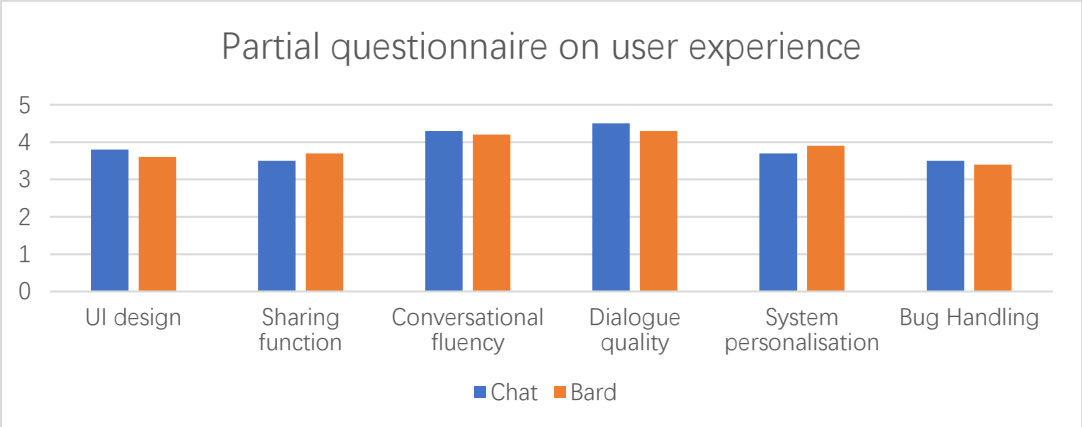
			Standardizer[a]	Point Estimate	95% Confidence Interval	
					Lower	Upper
Pair 1	A1 - B1	Cohen's d	1.581	1.054	.207	1.862
		Hedges' correction	1.752	.952	.187	1.680
Pair 2	A2 - B2	Cohen's d	2.404	1.063	.213	1.873
		Hedges' correction	2.663	.960	.193	1.691
Pair 3	A3 - B3	Cohen's d	2.121	.943	.128	1.719
		Hedges' correction	2.350	.851	.115	1.552
Pair 4	A4 - B4	Cohen's d	2.007	.775	.005	1.510
		Hedges' correction	2.223	.700	.004	1.363
Pair 5	A5 - B5	Cohen's d	1.787	1.243	.337	2.109
		Hedges' correction	1.980	1.122	.304	1.904
Pair 6	A6 - B6	Cohen's d	2.539	.700	-.052	1.419
		Hedges' correction	2.812	.632	-.047	1.281
Pair 7	A7 - B7	Cohen's d	1.716	.712	-.043	1.433
		Hedges' correction	1.901	.643	-.039	1.294
Pair 8	A8 - B8	Cohen's d	2.500	.667	-.078	1.378
		Hedges' correction	2.769	.602	-.070	1.244
Pair 9	A9 - B9	Cohen's d	2.205	.504	-.207	1.187
		Hedges' correction	2.442	.455	-.186	1.072
Pair 10	A10 - B10	Cohen's d	2.958	1.127	.258	1.956
		Hedges' correction	3.277	1.017	.233	1.766

a. The denominator used in estimating the effect sizes.
Cohen's d uses the sample standard deviation of the mean difference.
Hedges' correction uses the sample standard deviation of the mean difference, plus a correction factor.

Figure 4

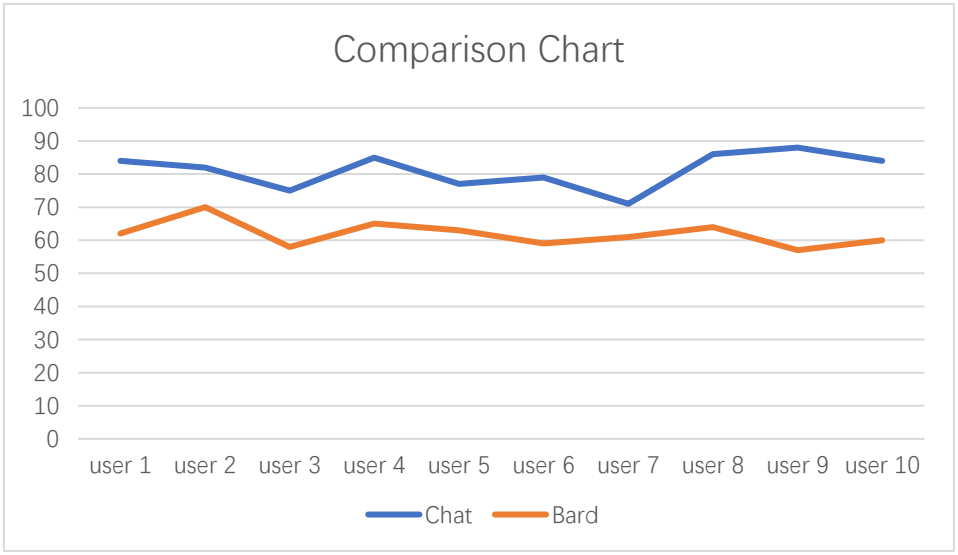
In conjunction with the qualitative feedback, we employed the SUS (System Usability Scale) survey, a standardized tool, to quantitatively evaluate users' perceptions of the usability of ChatGPT and Bard. This survey delves into the ease of use and user satisfaction metrics, offering a standardized scale to quantify the user experience (Law, Schaik & Roto 2014). The SUS questionnaire scores encapsulate user assessments across various facets of the tool, such as the user interface's ease of use, operational simplicity, and overall satisfaction levels.

By using A/B testing and qualitative feedback from SUS surveys, we gained a holistic view of users' perceptions and experiences with both ChatGPT and Bard. This comprehensive evaluation approach aims to uncover both nuanced qualitative insights and standardized quantitative measurements, contributing to a well-rounded understanding of these tools' usability and user satisfaction (Svensson 2023).



Score on the questionnaire: This scale goes from one to five with one meaning "strongly disagree" and five standing for "strongly agree". Though, half of the questions need to have their point deducted when they are used to calculate the score. This adjustment has the reason that the SUS focuses on the user's feeling ease the system is, while some questions are worded oppositely from others. Finally, the total of the ten responses was multiplied by 2.5 to give an overall score.

The following are the user feedback data from the A/B test and the System Usability Scale (SUS) questionnaire scores:



Results of data

This research uses a paired sample design to conduct a comparative analysis of the performance of two artificial intelligence conversation tools, ChatGPT and Bard. We observed that ChatGPT received higher

user ratings in all ten sets of comparisons, with average scores ranging from 5.67 to 9.00, while Bard's scores varied from 5.56 to 8.67, indicating that ChatGPT is generally higher in terms of user ratings than Bard. However, the range of standard deviations of the ratings indicates the variability that exists in user ratings. When conducting a paired samples t-test, we found that pairs 1, 3, 5, 6 and 10 showed statistically significant differences, with two-tailed p-values below the 0.05 significance level. These results indicate that there is a significant difference between users' ratings of ChatGPT and their ratings of Bard in these specific pairs. We quantified the practical significance of the differences between the two tools. In pairs 2, 4, 6 and 8, these larger effect sizes emphasise the advantage of ChatGPT over Bard in practical applications.

Based on the collected data and feedback from users, ChatGPT has garnered praise on multiple accounts. Users have commended its straightforward and user-friendly interface design, effortless stream of interactions, and organic responses. ChatGPT has achieved an average score of 67.2 on the System Usability Scale (SUS), which clearly indicates a high level of user appreciation for its comprehensive ease of use and usefulness. However, despite its diverse and comprehensive features, users tend to report that the Bard's intricate navigation design negatively impacts the overall user experience. With an average SUS score of 62.9, slightly lower than that of ChatGPT, the Bard has an inferior usability rating.In contrast, ChatGPT has received positive feedback for its interface design and user experience. Although the Bard offers rich features, its complex navigation is a significant obstacle to an enjoyable user experience. These findings serve as a reference for prospective enhancements of ChatGPT and Bard.

Although ChatGPT displays a superior competitive edge in this evaluation, the contrasting features of both models offer diverse options for users across different domains (Hannafin Land & Oliver 2013). ChatGPT and Bard are anticipated to persist in refining and enhancing their systems to cater to a diverse group of users and deliver a more comprehensive and gratifying human-computer interaction experience.

Conclusion:

This research comprehensively evaluated two AI conversational tools, ChatGPT and Bard. Through A/B testing and the SUS survey, we gained insight into the performance of the two tools in terms of user experience, interface design and interaction flow. The results show that ChatGPT generally scores higher than Bard in terms of user ratings, but the variability in ratings indicates a difference in user experience. The paired sample T-test showed significant differences between ChatGPT and Bard in certain situations, and the effect size analysis further highlighted the advantages of ChatGPT in practical applications.

ChatGPT was praised by users for its interface design, interactive fluency and natural responsiveness, and received a high average SUS score, indicating a high level of user approval for its ease of use and utility. In contrast, the Bard, although rich in features, had a negative impact on the user experience due to its complex navigation design, and its average SUS score was slightly lower than that of ChatGPT, indicating poor usability.

ChatGPT has garnered favorable reviews regarding interface design, interaction flow, and user satisfaction. The uncomplicated interface design and interactive style make it easier for users to comprehend and use. By contrast, the Bard's functional richness was acknowledged but posed a significant challenge to the user experience due to its intricate navigation architecture.

To enhance the Bard's complex navigation, simplifying the steps or presenting a more intuitive user guide should be considered. The layout of the ChatGPT interface and the clarity of its functions could be improved to enhance the user experience.

Overall, ChatGPT showed a competitive advantage in this evaluation, but the features of both give users a variety of options in different areas. Despite ChatGPT displaying a stronger competitive advantage in this assessment, both platforms' traits offer a diverse range of options across various fields Both ChatGPT and Bard are expected to continue to improve their systems to meet the needs of different user groups, opening up more possibilities for a more comprehensive and satisfying human-computer interaction experience. This research provides important references and guidance for the future development of AI dialogue tools.

Reference:

Lazar, J., Feng, J. H., & Hochheiser, H. (2017). *Research methods in human-computer interaction*. Morgan Kaufmann.

Mattas, P. S. (2023). ChatGPT: A Study of AI Language Processing and its Implications. *Journal homepage: www. ijrpr. com ISSN*, *2582*, 7421.

Tidwell, J. (2010). *Designing interfaces: Patterns for effective interaction design*. " O'Reilly Media, Inc.".

Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*.

Hannafin, M., Land, S., & Oliver, K. (2013). Open learning environments: Foundations, methods, and models. In *Instructional-design theories and models* (pp. 115-140). Routledge.

Law, E. L. C., Van Schaik, P., & Roto, V. (2014). Attitudes towards user experience (UX) measurement. *International Journal of Human-Computer Studies*, *72*(6), 526-541.

Khennouche, F., Elmir, Y., Djebari, N., Himeur, Y., & Amira, A. (2023). Revolutionizing customer interactions: Insights and challenges in deploying chatgpt and generative chatbots for faqs. *arXiv preprint arXiv:2311.09976*.

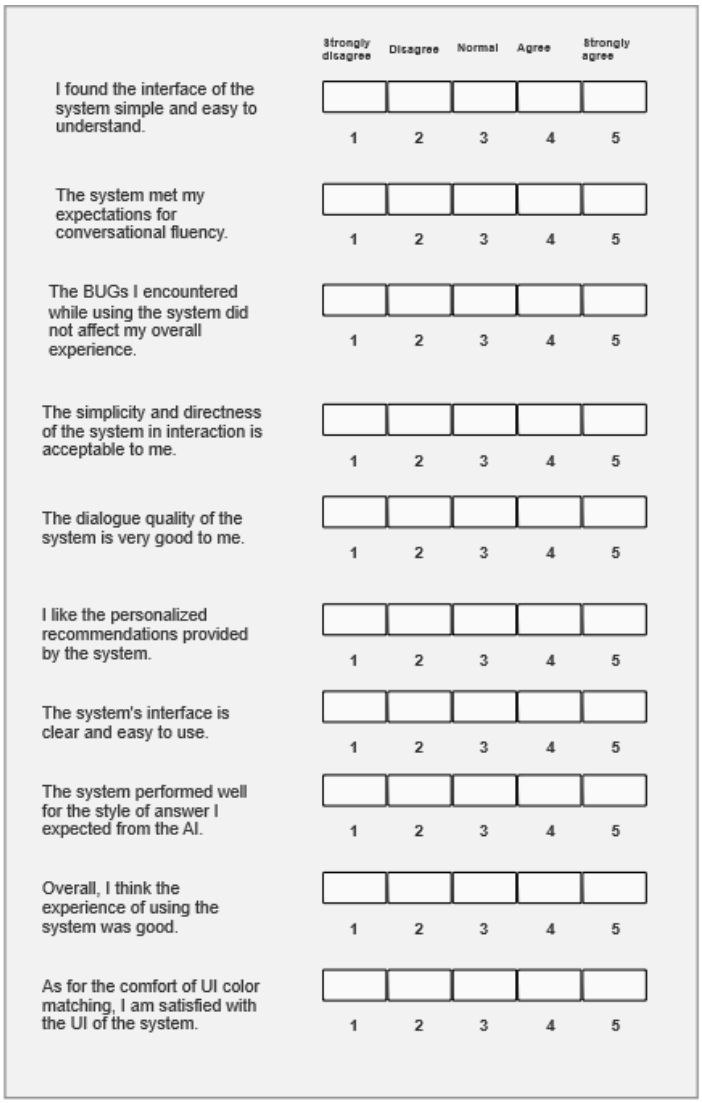
Svensson, R. (2023). Contextualizing Customer Feedback: A Research-through-Design Approach-Alternative Approaches and Dialogical Engagement in Survey Design.

Ren, X., Silpasuwanchai, C., & Cahill, J. (2019). Human-engaged computing: the future of human–computer interaction. *CCF transactions on pervasive computing and interaction*, *1*, 47-68.

Karat, J., & Karat, C. M. (2003). The evolution of user-centered focus in the human-computer interaction field. *IBM Systems Journal*, *42*(4), 532-541.

Appendix:

SUS Questionnaire:



	Strongly disagree	Disagree	Normal	Agree	Strongly agree
I found the interface of the system simple and easy to understand.	1	2	3	4	5
The system met my expectations for conversational fluency.	1	2	3	4	5
The BUGs I encountered while using the system did not affect my overall experience.	1	2	3	4	5
The simplicity and directness of the system in interaction is acceptable to me.	1	2	3	4	5
The dialogue quality of the system is very good to me.	1	2	3	4	5
I like the personalized recommendations provided by the system.	1	2	3	4	5
The system's interface is clear and easy to use.	1	2	3	4	5
The system performed well for the style of answer I expected from the AI.	1	2	3	4	5
Overall, I think the experience of using the system was good.	1	2	3	4	5
As for the comfort of UI color matching, I am satisfied with the UI of the system.	1	2	3	4	5

The formula for paired samples t-test:

$$t = \frac{m}{s/\sqrt{n}}$$

m: the mean of the sample differences

n: the sample size

s: the standard deviation of the sample difference, and the degree of freedom df is n-1

We can calculate the t test statistic (|t|) for the degrees of freedom (df) and compare its P value at df=n-1 by querying the t distribution table.

If the p-value is less than or equal to the significance level of 0.05, we can reject the null hypothesis and accept the alternative hypothesis. In other words, we conclude that the two paired samples are significantly different.s

ChatGPT feedback data:		SUS score:
User 1	Simple and easy to understand interface, smooth interaction.	68
User 2	Appreciate the AI's natural replies, but some of the conversation lacks coherence.	72
User 3	There were a few bugs encountered during use, but the overall experience was pretty good.	65
User 4	The interaction is simple and straightforward, but sometimes the response time is a little long.	70
User 5	Dialogue quality is good, but the interface layout is a bit cramped.	67
User 6	Love the personalised recommendation feature, but the positioning of some functionality buttons needs clarification.	69
User 7	Liked the personalised recommendations feature, but the placement of some of the function buttons was not clear.	71
User 8	Liked AI's style of reply, but misinterpreted in some cases.	66
User 9	The overall experience of using it is good, but individual features are not intuitive enough.	68
User 10	The dialogue quality is good, but the UI colour scheme is not comfortable enough.	64

Bard feedback data:		SUS scores:
User 1	The functional diversity is impressive, but the interface layout is slightly complex.	62
User 2	Navigation is easy to get lost in the options, but the interactivity is great.	70
User 3	The quality of dialogue and content is excellent, but the navigation is not intuitive.	58
User 4	Functionality is good, but certain features need better user guidance.	65
User 5	The UI is clean, but the location of certain features is unclear.	63
User 6	Informative, but slightly complicated to operate.	59
User 7	The interface has a good colour scheme, but certain features need improvement.	61
User 8	Interactions are smooth, but individual functions are not clear enough.	64
User 9	The overall experience is average and needs to be improved with user guidance.	57
User 10	Some of the features are not intuitive enough, but the dialogue quality is good.	60