AI Search Showdown: Uncovering User Experience Insights in Google Bard and Microsoft Bing

Ahmet Yasin Aytar

Atra Zeynep Bahçeci

Halil İbrahim Ergül

Serhat Demirkıran

Abstract

This paper presents a comparative study of user experience (UX) with two leading AI-based search engines, Google Bard and Microsoft Bing, across a variety of everyday and domainspecific tasks. Through quantitative metrics such as task completion time, accuracy, and relevancy scores, along with qualitative user feedback, we evaluated the performance of each search engine in terms of usability. The study also utilized the System Usability Scale (SUS) for a standardized assessment of usability and an Ordinary Least Squares (OLS) regression model to understand the impact of demographic factors on user satisfaction. Our findings suggest that while Google Bard was generally preferred and perceived as relatively successful compared to Microsoft Bing, both search engines exhibited distinct strengths and weaknesses in specific areas. The study's insights highlight the importance of nuanced AI integration and interface design to enhance the efficacy and user-friendliness of AI-driven search tools. This research contributes to the broader understanding of AI-UX interaction and provides actionable guidance for developers and designers to refine AI-based search technologies.

1.1) Background of the Study

The “AI Boom” of the last decade introduced new Generative Artificial Intelligence (GenAI) tools to many domains; from education to health, insurance to banking. Surely, the search engine industry with over a 160 billion USD market size was no exception (Business Research Insights, 2023). Applications such as ChatGPT, Bing Chat, Google Bard, and Opera Aria have over 180 million users despite being released in the last 2 years (Tong, 2023). All these search engines and tools cater to a wide variety of users, enhancing the importance of user experience.

A 2023 study compared ChatGPT and Google in terms of productivity and user satisfaction. Users were given 3 tasks; one for finding a fact, one for finding relevant websites, and one for assessing the truth of given statements. Productivity, measured by task completion time, was 158% higher with ChatGPT. Satisfaction, measured via a questionnaire on information quality, ease of use, usefulness, enjoyment, and satisfaction, was again higher with ChatGPT (Xu, et. al). This study confirmed that users prefer AI-based engines over regular search engines, and thus hints at the shift towards AI-based platforms.

Another paper compared ChatGPT, Google BARD, and Microsoft Bing regarding Natural Language Processing (NLP), Machine Learning (ML), and UX. The participants, Croatian higher education students, were asked to complete a series of tasks to create a presentation on various subjects. Quantitative and qualitative data was collected on variables accuracy, response

time, relevance, user satisfaction, and engagement. The findings showed that all search tools excelled at different aspects; with ChatGPT with the most accurate and relevant tool, Google BARD with the fastest response time, and Microsoft Bing with the highest user satisfaction and engagement (2023, Bhardwaz & Kumar).

A 2023 study conducted at the Nielsen Norman Group investigated the user preferences of ChatGPT, Google BARD, and Microsoft Bing in terms of helpfulness and trustiness. Users found Bing Chat the least helpful and the least trustworthy. Interestingly, the low user ratings of Microsoft Bing were not only rooted in an NLP issue of giving broad answers but also in the UX issue of not displaying reference links in an easily accessible manner (Liu, et. al).

Researchers Skujve et. al conducted a questionnaire on early users of ChatGPT. Their findings show that the user experience is not only impacted by pragmatic attributes such as efficiency, relevancy, and accuracy; it is also greatly impacted by hedonistic attributes like creative impact, entertainment, and surprise (2023). These findings once again highlight that GenAI tools are not only a research area for NLP engineers, but also for UX researchers.

Other than the related work mentioned above that investigates the UX of AI-based search engines for everyday tasks; there are various papers examining the applicability and UX of these tools in specific domains, such as education (Dao, 2023), and psychology (Kumar, et. al, 2022). In this study, the user preferences between Google BARD and Microsoft Bing will be researched, for both everyday tasks and domain-specific inquiries. The attributes of these tools affecting the user experience will be investigated. This study is important because it provides a comprehensive analysis of user experience (UX) by comparing two leading AI-based search engines across a range of everyday and domain-specific tasks. Unlike prior research, which has largely focused on individual aspects of UX or specific domains, this study encompasses a broader spectrum of usability factors, including task completion time, accuracy, relevance, and overall user satisfaction. By investigating the nuanced preferences and behaviors of users when interacting with these search engines, the study sheds light on the practical efficacy of AI models embedded within them.

2.1) Methodology

This research aims to answer the following questions:

i. How do the attributes of response accuracy, speed, and resource clarity influence user satisfaction and preference when utilizing AI-based search engines for both generic and domain-specific queries in Google Bard and Microsoft Bing?
ii. To what extent does the integration of AI search engines of Bard and Bing with existing services and tools, such as code explanation and mapping features, impact the overall user experience in terms of efficiency and task completion?

Though no data with a valid source is available for the user demographics of Microsoft Bing Chat, Google BARD states that its users are predominantly between the ages of 14-29; and relatively balanced in terms of gender with a slight skew towards male users (2024). Hence, the following user persona is created considering the actual user demographics and the aims of this study:

Name: Ahmet Yılmaz

Occupation: Master's Student & Software Engineer

Education: BSc in Computer Engineering, Currently pursuing MSc in Computer Science

Age: 24

Profile: Ahmet is a dedicated master's student at Sabanci University and is simultaneously working remotely as a software engineer. His consistent success throughout his academic and professional career is fueled by his hard work and genuine interest in computer science/technology. He is fluent in English, and a native Turkish speaker. He is of Turkish origin and is currently residing in Istanbul, Turkey.

Goals:

- To explore the world of GenAI tools, which he believes he can utilize both as a student and a professional.
- To gain knowledge on AI-based search engines; their capabilities, interfaces, integrations, and validity.
- To contribute to the field of UX on GenAI tools and also to experience how a user test is performed.

Needs:

- AI-based search engine interfaces that are easy to discover and learn.
- AI-based search engines that can save him time and energy for his academic and professional projects.
- AI-based search engines that he can use for both everyday, general needs, and for coding support.

Technical Background and Workplace: Ahmet has a very strong foundation in computer science and a broad understanding of GenAI. He spends most of his time on his MacBook Pro, for his job, his education, and also entertainment. He prefers the Google Chrome and the Mozilla Firefox web browsers and uses both interchangeably. As a software engineer, he understands the importance of UX and interface design; and has taken relevant elective courses during his undergraduate education.

General Background: Ahmet is a very hard-working and curious individual. He is a self-starter. He finds the recent tools built on Large Language Models (LLM) very promising. Though he is not an AI Engineer, he has researched and studied the fundamentals of these models. He frequently uses ChatGPT and has a ChatGPT 4.0 membership that allows him to do web searches. However, he has not tried out the Google BARD or Microsoft Bing on their interfaces.

Archetype: Ahmet can be considered the epitome of a *"Straight-A Student"* and *"Tech Enthusiast"*. He finds the new technological advancements very interesting and takes time out of his busy schedule to keep up-to-date with the current technologies and see how he can leverage them to increase his productivity at work and school.

Experience Goals: Ahmet expects the AI-based search engines to provide fast and accurate answers. He anticipates that these tools can perform successfully for his daily inquiries and for coding support. He also hopes to have an entertaining experience and is curious to see if these tools can generate any surprising/fun answers.

Brand-Relationship: He has a positive perception of both Google and Microsoft, he uses various tools such as Google Drive, Google Colab, Microsoft Office, et cetera daily. He is not biased towards one, as he understands how the performance of a tool from a brand might not reflect the other.

User Persona

11 volunteers participated in the study, with attributes in line with the user persona. Demographic information of users was collected via a pre-test questionnaire, listed as *Pre-Test Questionnaire* in the Appendix. As suggested by Tankala, the users had the "Prefer not to Say" option for questions that may be sensitive, and the ages were collected via interval selections, not direct entry (2022). Data on the user's computer skills, previous experience, and interest in AI-based search engines were also collected. The results of the questionnaire are given below in the appendix as *Pre-Test Questionnaire Results,* and summarized in Table 1.

Gender	Female (45.5%)		Male (54.5%)		Prefer not to say (0%)
Age	20-29 (90.9%)		30-39 (9.1%)		Prefer not to say (0%)
Education Level	High School (0%)	Bachelor's Degree (45.5%)	Master's Degree (45.5%)	Doctorate's Degree (9.1%)	Prefer not to say (0%)
Occupation	Student / Intern (54.5%)	Engineer (18.2%)	Data Scientist (9.1%)	Researcher (9.1%)	Financial Analyst (9.1%)
Nationality	Syria (9.1%)		Turkey (100%)		Unanswered (0%)
Residency	Turkey (100%)				Unanswered (0%)
Overall Computer Skills (1-5 Scale)	1 (0%)	2 (9.1%)	3 (27.3%)	2 (18.2%)	5 (45.5%)
Previous Experience with AI-Based Search Engines	Yes (54.5%)		No (45.5%)		
Opinion on AI-Based Search Engines (1-5 Scale)	1 (0%)	2 (0%)	3 (9.1%)	4 (45.5%)	5 (45.5%)
Interest in AI-Based Search Engines (1-5 Scale)	1 (0%)	2 (9.1%)	3 (18.2%)	4 (27.3%)	5 (45.5%)
Perceived Visual Ability (1-5 Scale)	1 (0%)	2 (9.1%)	3 (9.1%)	4 (36.4%)	5 (45.5%)
Physical or Cognitive Impairments	Yes (0%)		No (100%)		

Table 1: Participant Demographics & Skills

Overall, as seen in Table 1, the participants are young, highly educated with good computer skills. All have a relatively high interest in AI-Based Search Engines, and about half have previous experience with them. Most are educated/work in technology-related fields, but there is still enough diversity among the participants in terms of their field of work without compromising the technical skills required for the test. Additionally, none of the participants have physical or cognitive obstacles that may skew the results of the study.

The user tests were performed on Google Meets, as it allows the testers to see the participants' screen without having to hover over them or disturb them in a physical environment.

Before the test, participants were informed that they would perform 5 tasks on Google BARD and Microsoft Bing sequentially. The participants were guided to the related platforms via a link and were given time to shortly discover, or "get the feel" of the interfaces.

Users were required to complete 5 tasks as the following: i. **Fact**

Retrieval with queries:

"Who is the first female pilot in Turkey?"

"Who wrote the novel İnce Memed?"

"Who directed the 2011 film Once Upon a Time in Anatolia?"

After the search engine generated answers, participants were encouraged to check out the links provided, if any, to assess the accuracy and reliability of the answer. This task is adapted from the 2023 study on the usability of ChatGPT and Google Search (Xu, et al.).

ii. **Current News Summarization** with the query:

"Summarize the top 5 today's news for Turkey for today?"

Once the answer was generated, participants were instructed to check the reference links and to investigate whether or not the news was up-to-date. A recent study at the Nielsen-Norman Group concluded that the link placements and formats greatly affect user satisfaction for AIbased search engines, and hence specific importance has been put on the links for this task (Liu, et al., 2023). iii. **Local Cuisine Recommendations** with the query:

"Recommend me 5 best local dishes in Istanbul"

For this task, the users were encouraged to think aloud about the dish descriptions, the places recommended to experience these dishes, and the relevancy and diversity of the answers. This task was inspired by multiple studies and created to measure the extent of the answers (Liu, et. al, 2023), the entertainment/surprise aspect (Skujve, et. al, 2023), and relevance (Bhardwas, et. al, 2023).

iv. **3-Day Travel Itinerary for Rome** with the query:

"Write to me 3-day travel itinerary for the city of Rome"

The participants focused on the day-wise breakdown of activities, sightseeing spots, local transportation information, and dining and accommodation suggestions. They were directed to go to the reference links generated, and assess the reliability of the answer. Once again, this task was designed to test the element of surprise, the extent of the answers, reliability, and relevance).

v. **Coding Quality** with the queries:

"Write the Python code for the following task:

Given an integer array nums, return an array answer such that answer[i] is equal to the product of all the elements of nums except nums[i]. You must write an algorithm that runs in O(n) time and without using the division operation.

Example:

Input: nums = [1,2,3,4]

Output: [24,12,8,6]"

"Re-write the code using the division operation"

For this task, the users commented on the accuracy of the code generated, how it was explained, how relevant were the function and variable names, and the sharing options.

Standardized queries were used, meaning all the participants used the same queries provided by the tester. This is to ensure that the assessment is on the usability of the tools, not their performances as LLMs. 6 participants started with Microsoft Bing, and 5 started with Google BARD. The tests were not recorded as most users refused to participate in a recorded test. The data was collected in the following forms:

- Task completion times
- Notes taken of participant comments during the test.
- Standard Usability Scale (SUS) participants completed for each platform after the test.
- Post-Test Questionnaire

The forms completed after the test are given in the appendix under *SUS* and Questionnaire. The Post-Test Questionnaire consisted mostly of 1-5 scale questions and collected overall user preferences, along with task-based data. The questionnaire measured the following elements for each task:

- Task Completion
- Response Time
- Accuracy
- Trustiness
- Element of Surprise
- Link Placement and Formatting
- Extent of the Answers

The participants were encouraged to complete the forms without dwelling on their answers as suggested by Brooke (1996).

3) Results / Analysis of Data

In this section, we show the outcomes of our usability testing, derived from analyzing subjects' task performances and responses to our post-experiment questionnaires.

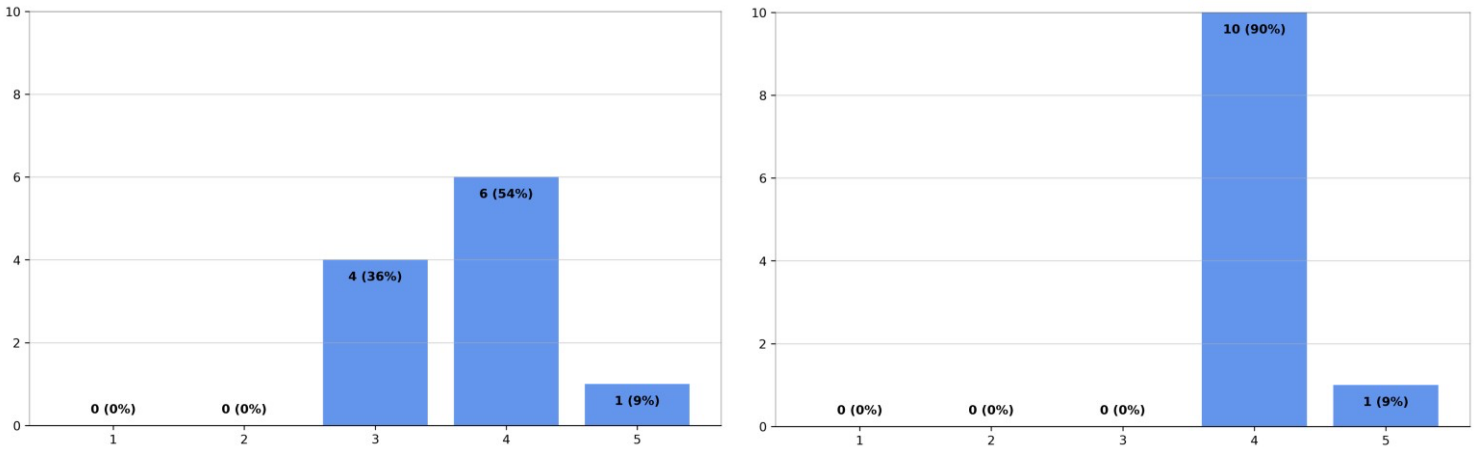

Figure 1: Overall Level of Satisfactions with Bing (Left) and Bard (Right)

Upon examining the results of our usability testing, which aimed to compare two renowned AIbased search engines—Google Bard and Microsoft Bing—the data depicted in the provided visualizations reveals notable differences in user satisfaction and preference.

The bar graphs in **Figure 1**, illustrating the overall level of satisfaction with Microsoft Bing and Google Bard respectively, convey distinct user experiences with the two search engines. In the case of Microsoft Bing, a majority of users rated their satisfaction at a level 4, suggesting a commendably high level of satisfaction, albeit with some room for improvement. A significant proportion of participants rated their satisfaction at level 3, indicating moderate satisfaction with the service. Notably, a minority of users expressed the highest level of satisfaction by rating it at level 5. The absence of satisfaction ratings at levels 1 and 2 for Microsoft Bing indicates that the search engine did not elicit low levels of satisfaction among the participants in our study.

In stark contrast, Google Bard received an overwhelmingly positive reception, with a remarkable 90% of users rating their satisfaction at level of 4. A small contingent of users rated their satisfaction at level 5, which denotes a high level of satisfaction. The complete lack of ratings at levels 1, 2, and 3 for Google Bard underscores the search engine's exceptional performance in user satisfaction within the context of the tasks assigned during the usability testing. After user tests were completed, we asked subjects "Which engine do you prefer, Google Bard or Microsoft Bing?". Results showed that participants had a more satisfying experience with Google Bard. In **Figure 2**, it can be seen that %63 percent of respondents preferred Google Bard over Microsoft Bing.

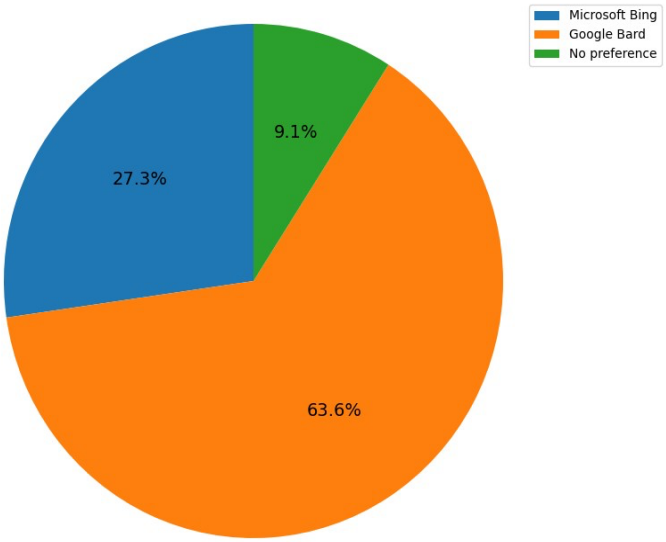

Figure 2: Search Engine Preference

A substantial majority of 63.6% manifested a preference for Google Bard, showing the high satisfaction ratings observed in the bar graph. Conversely, Microsoft Bing was preferred by 27.3% of users, while a minority of 9.1% expressed no preference between the two. The analysis of these results indicates a pronounced preference and higher levels of user satisfaction with Google Bard over Microsoft Bing. The inclination towards Google Bard is particularly evident in the context of the five tasks designed for the experiments, which encompassed fact retrieval, news summarization, local cuisine recommendations, travel itinerary composition for Rome, and domain-specific coding quality.

A. Quantitative Metrics: Completion Time, Accuracy, and Relevancy

To understand further the participants' preference of one over the other, we also asked questions to have an idea about what they think and how they perceive the two search engines' performances in more quantitative terms: completion time, accuracy, and relevancy. For each task, participants were instructed to provide feedback regarding different aspects, including response time, accuracy, relevancy, and more. Participants assigned likert-scale scores ranging from 1 (indicating strong disagreement) to 5 (indicating strong agreement) for their responses. We then computed the average scores for accuracy, relevancy, and response time across all tasks for both Microsoft Bing and Google Bard. Below, we have corresponding figures for each attribute. We note that for accuracy, and relevancy scores we did not include comparison for Task 5 (coding). The rationale behind is that because Task 5 was more domain-specific compared to other tasks and it would be misleading to measure this task along with others. Therefore, we prefer to analyze users' responses for both Bing and Bard by focusing on whether the explanation of the code was clear and helpful without needing a second prompt.

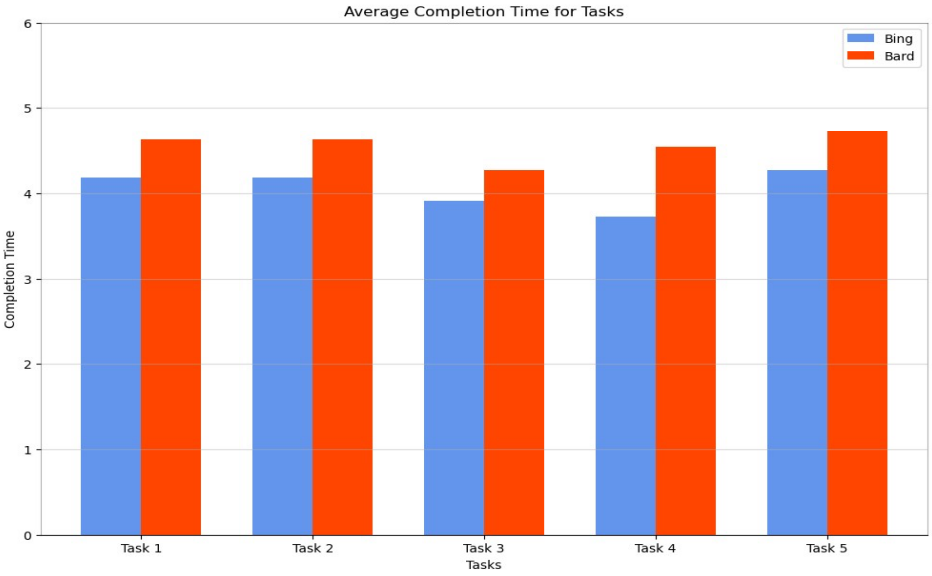

Figure 3: Completion Time for Bing and Bard

The data presented in the **Figure 3** indicates that for every task, participants observed quicker response times with Google Bard compared to Microsoft Bing. Informed by the literature, we concluded that the completion time for tasks is a significant indicator of the efficiency of the search engines. The data indicates that Google Bard tended to enable faster completion of all tasks overall, suggesting a more streamlined search and retrieval process. Especially for Task 4 Bard proves to be more efficient.

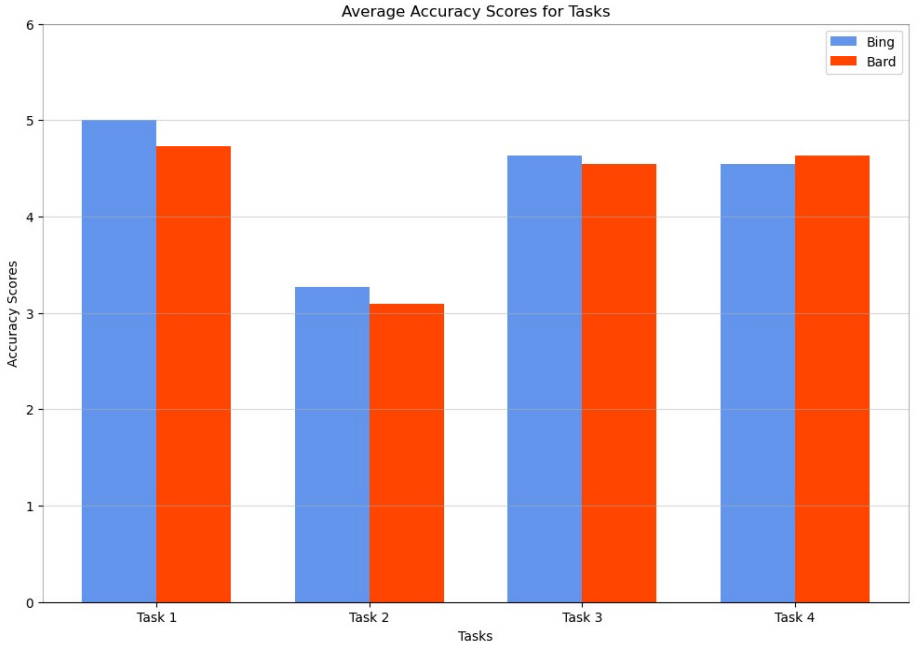

Figure 4: Accuracy Scores for Bing and Bard

Upon closer inspection of **Figure 4**, which delineates the average accuracy scores for tasks conducted using Microsoft Bing and Google Bard, it is apparent that the performance of both search engines is closely matched across all tasks. For Task 1, the accuracy scores are indeed very close, with Microsoft Bing and Google Bard demonstrating near parity. This task, centered

on retrieving a fact—specifically, identifying the first female pilot in Turkey—requires precision and a reliable knowledge base. The comparable scores suggest that both search engines have a robust database for such factual queries and are equally proficient in accessing this information. Task 2, which involves summarizing current news, again shows a marginal difference in the accuracy scores between the two search engines. The slight variance indicates that both Bing and Bard have effective mechanisms for filtering and condensing news content, although one might be slightly more refined than the other.

Similarly, Task 3's results reveal a very tight race between the two search engines in recommending local dishes in Istanbul. This suggests that both Bing and Bard are wellequipped to parse local information and user reviews to generate relevant satisfactory suggestions. In Task 4, the task of crafting a 3-day travel itinerary for Rome is met with virtually identical accuracy by both search engines. This indicates a comprehensive ability to access and organize travel-related data, such as points of interest, local transportation options, and events, which is essential for itinerary planning.

The close accuracy scores across all tasks in Figure 4 suggest that Microsoft Bing and Google Bard are highly competitive in terms of accuracy. For the users and tasks involved in this study, neither search engine consistently outperformed the other to a significant degree.

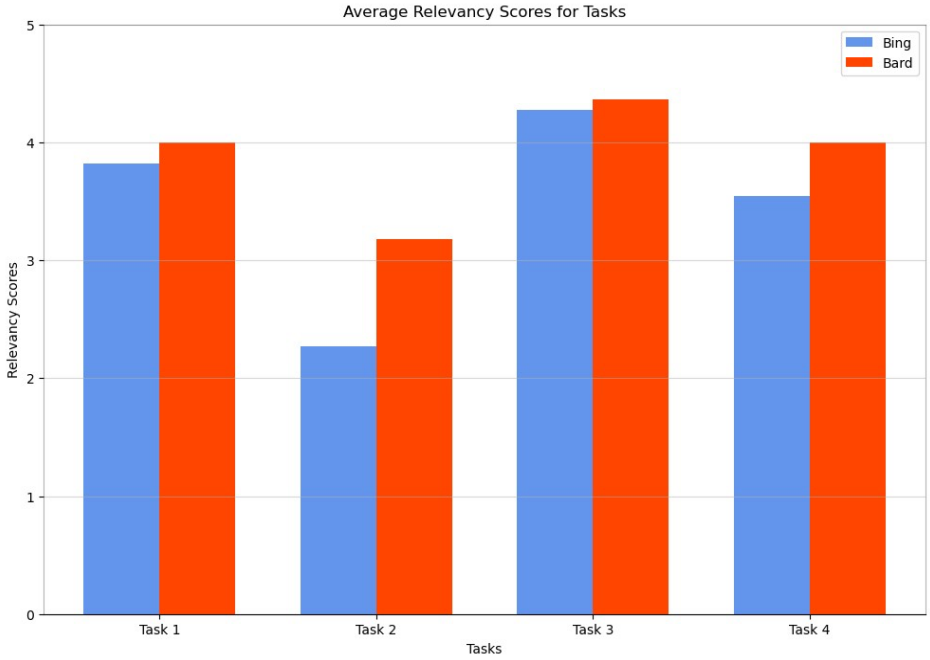

Figure 5: Relevancy Scores for Bing and Bard

Figure 5 presents the average relevancy scores attributed to Microsoft Bing and Google Bard across four distinct tasks. Relevancy, in this context, measures the extent to which the search engine results matched the users' information needs and query intent. For Task 1, Microsoft Bing's relevancy score is marginally lower than that of Google Bard, indicating that users found Bard's responses slightly more aligned with their query about the first female pilot in Turkey. The higher relevancy for Bard could suggest that its search results were more directly connected to the query or that the information presented was perceived more informative by the users.

Task 2 shows a significant difference, with Bard leading by a more substantial margin. This task involved summarizing the top news for the day, which likely required the integration of various information sources and a coherent presentation. Bard's higher relevancy score suggests that its summaries were more on point, probably providing a clearer, more concise, or more comprehensive synthesis of the day's news in a manner that users found more useful. In Task 3, which asked for recommendations on the best local dishes in Istanbul, both search engines received nearly identical relevancy scores, with Bard having a slight edge. This task demanded an understanding of local culture and cuisine, and the equivalent scores imply that both search engines are well-equipped to tap into local content and user reviews to generate relevant suggestions.

Task 4, the creation of a 3-day travel itinerary for Rome, again saw comparable relevancy scores for both Bing and Bard, with Bard slightly ahead. This task required the retrieval of a diverse set of data, including points of interest, local events, and transportation options. The close scores suggest that both search engines are similarly capable of compiling relevant travel information, although Bard may have a slight advantage in how it presents this information or in the completeness of the data provided. In a nutshell, the data from **Figure 5** illustrates that while there are some fluctuations in the relevancy scores between tasks, the performance of the two search engines is closely competitive, with Google Bard generally perceived as slightly more relevant across the tasks.

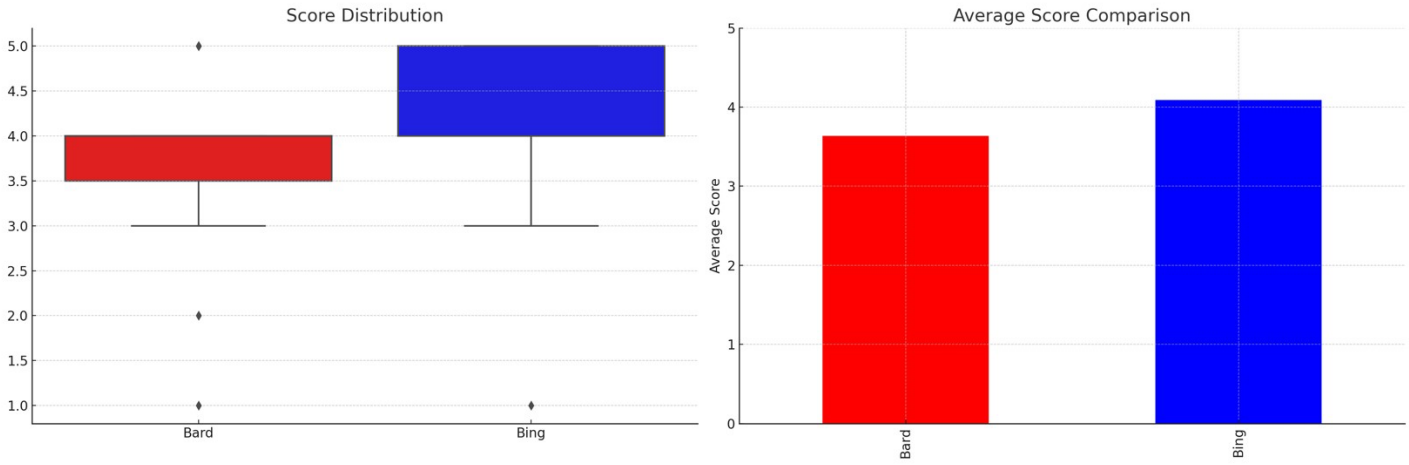

Figure 6: Score Comparison for Task 5

In the assessment of Task 5, which entailed a domain-specific coding task, the comparative analysis of participant-generated scores was conducted for Bard and Bing. The score distribution, visualized via a box-and-whisker plot (**Figure 6**, left), indicates a statistically lower median score for the Bard, represented by a red box, in contrast to the Bing, denoted by a blue box. The interquartile range for Bard is more constricted, suggesting a higher consistency in participant scoring; however, the presence of outliers, depicted by individual points below the primary quartile box, reveals instances of significantly lower scores that deviate from the normative scoring pattern. The Bing score distribution is more dispersed, implying a greater variability in the scores assigned by participants within this group.

The average score comparison, illustrated in a bar graph (**Figure 6**, right), corroborates the aforementioned distribution patterns, with Bard obtaining a lower mean score relative to Bing. This discrepancy in average scores further underscores the differential evaluative tendencies of the participants when engaged in the scoring of Task 5.

In summation, the findings suggest that participants evaluating the domain-specific coding task attributed a higher range and average of scores to the Bing compared to the Bard. The lower score of Bard in this domain-specific task could be attributed to one of our findings in verbal feedback we obtained from the participants. Some of them stated that Bard's explanations were less satisfactory in code-related answers, and it had a missing clarity in the sources or links used to generate those answers. In terms of Bard, they were dissatisfied with the accuracy of information and the lack of example usage in Python code.

B. SUS Scores Distributions

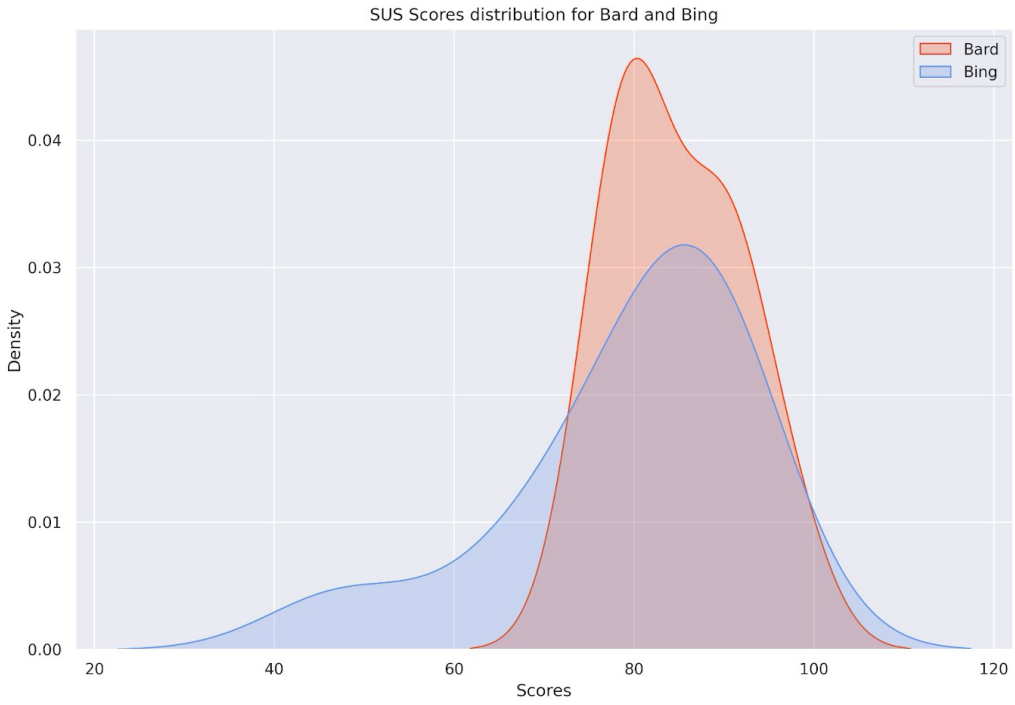

Figure 7: SUS Scores Distributions

The SUS is a reliable tool for measuring the usability of various systems, including software interfaces, providing a standardized scale that ranges from 0 to 100, where higher scores denote better usability. The distribution curves in **Figure 7** for both search engines reveal insights into their perceived usability. The curve for Google Bard (represented in red) displays a distribution that is skewed towards higher SUS scores, peaking just below the score of 80 before tapering

off. This peak indicates that a majority of users rated Bard with high usability scores, suggesting that they found the interface to be user-friendly, efficient, and satisfying. Conversely, the distribution for Microsoft Bing (depicted in blue) peaks at a lower SUS score, around the mid60s, and presents a broader spread across the SUS score range. This broader distribution indicates a greater variance in user responses regarding Bing's usability, with scores ranging from the lower end of the scale up through the high 70s. The peak at the lower score suggests that, on average, users found Bing to be less usable compared to Bard.

Overall, the graph signifies that while both systems were rated as usable, Google Bard tends to be perceived as more usable than Microsoft Bing. The distribution of scores for Bard is concentrated in the higher ranges, showing a consensus among users regarding its favorable usability. In contrast, the wider spread of Bing's scores suggests more varied user experiences, which could point to inconsistencies in the user interface or interactions that some users may find less intuitive or efficient.

4) Discussion and Conclusion

This study aimed to evaluate user preferences and experiences with Google BARD and Microsoft Bing in the context of various tasks, including fact retrieval, current news summarization, local cuisine recommendations, travel itinerary creation, and domain-specific coding quality assessment. Google Bard seems to be edging out Microsoft Bing in most tasks by a narrow margin and being generally favored by participants. This is interesting to observe even though Bing was found to be more useful and successful in terms of the last domain specific coding task (Task 5). However, to understand more what made our participants to choose for Bard and how their ideas for both of tools were shaped we also gathered verbal responses. This feedback reveals specific areas of dissatisfaction with each search engine.

For Google Bard, users expressed concern over the use of subjective language such as "legendary" or "epic," which could be seen as injecting opinion rather than maintaining objectivity in search results. A notable deficiency in Bard was the lack of satisfactory explanations in code-related answers and a missing clarity in the sources or links used to generate those answers. Users expected more from Bard's integration with Google's services, such as Maps, to enhance the relevance of recommendations. Lastly, users were dissatisfied with the accuracy of information and the lack of example usage in provided Python code, which could have implications for the perceived reliability and educational value of the engine.

Microsoft Bing faced criticism primarily for its slower response times and less detailed answers. Users were particularly critical of the inefficiency in handling the Rome itinerary question, noting slow responses and irrelevant links. The feedback mechanism on Bing was seen as lacking, as users couldn't easily revisit their previous inputs and found the user interface elements like the prompt box and answer toggles unintuitive. Moreover, users found the transition between search and copilot features to be potentially confusing. A recurrent theme in the feedback for Bing was the quality of links provided, with users encountering broken links or being redirected to Bing search results instead of direct answers. We believe this is interesting

because previous studies similarly show that the low user ratings of Microsoft Bing were not only rooted in an NLP issue of giving broad answers but also in the UX issue of not displaying reference links in an easily accessible manner (Liu, et. al). In contrast with this finding we observed that most of the users interpreted Google Bard as less capable of providing links in an easily accessible manner. Lastly, there was a sentiment that Bing's presentation of links as authoritative could mislead users, especially when the accuracy of those links was not verified.

In a study regarding comparison of Google Bard, ChatGPT, and Microsoft Bing, Bard has been found with the fastest response time relatively (Bhardwaz & Kumar). Our results regarding the completion time of Google Bard and Microsoft Bing are in align with this finding that response time of Google Bard outperforms Microsoft Bing from the perception of participants.

We also implemented two statistical significance tests (OLS regression for Bard and Bing) to explain the relationship between overall satisfaction (dependent variable) with general interest in the tool (denoted as interest in the table in Appendix [5]), trust to the tool, immersion, gender, education, and occupation. Although results do not reveal a statistically significant relationship between these variables, we interestingly found that the gender coefficient is 1.68, with a Pvalue of 0.044, which is significant at the 5% level (indicated by **). This suggests that being male is associated with a higher satisfaction with Bing, holding other variables constant. Other variables such as age, interest in the product, trust, immersion, level of education, and occupation do not appear to have a statistically significant impact. However, it is important to note that the significance of these results could be influenced by the sample size and the representativeness of the sample population. Additionally, further research could be conducted to explore other potential variables that may affect user satisfaction with search engines.

Lastly, based on our findings in SUS score distributions and overall score comparisons (completion time, accuracy, relevancy), we conclude that the superior satisfaction ratings for Google Bard suggest that it may have delivered more accurate, relevant, or user-friendly responses, potentially attributable to the search engine's AI model's capabilities, the design of the user interface, or the efficiency of the engine in parsing and responding to the user queries. For developers and designers, these results emphasize the importance of not only the functional capabilities of AI-based systems but also the overall user experience, which encompasses ease of use, learnability, and satisfaction. The findings could guide future enhancements, focusing on areas where users encountered usability issues with Bing, and reinforcing the strengths of Bard that contributed to its higher SUS scores.

References

Bhardwaz, S., & Kumar, J. (2023). An Extensive Comparative Analysis of Chatbot Technologies—ChatGPT, Google BARD and Microsoft Bing. *2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, 673–679. https://doi.org/10.1109/ICAAIC56838.2023.10140214

Brooke, J. (1996). SUS: A quick and dirty usability scale. In P. W. Jordan, B. Thomas, I. L. McCelland, & B. Weerdmeester (Eds.), Usability Evaluation in Industry (1st ed., pp. 189-194). CRC Press. https://doi.org/10.1201/9781498710411

Business Research Insights. (2023). *Global search engine market size, share report | Forecast from 2023 to 2031*. https://www.businessresearchinsights.com/market-reports/search-enginemarket-101546

Dao, X. Q. (2023). *Which Large Language Model should You Use in Vietnamese Education: ChatGPT, Bing Chat, or Bard?* (SSRN Scholarly Paper 4527476). https://doi.org/10.2139/ssrn.4527476

Google BARD (2024, January 16). *Google BARD's User Base.* Retrieved January 16, 2024, from https://g.co/bard/share/7a87b009c0a6

Kumar, H., Musabirov, I., Shi, J., Lauzon, A., Choy, K. K., Gross, O., Kulzhabayeva, D., & Williams, J. J. (2022). *Exploring The Design of Prompts For Applying GPT-3 based Chatbots: A Mental Wellbeing Case Study on Mechanical Turk* (arXiv:2209.11344). arXiv. https://doi.org/10.48550/arXiv.2209.11344

Liu, F., Budiu, R., Zhang, A., & Cionca, E. (2023, October 1). *ChatGPT, Bard, or Bing Chat? Differences Among 3 Generative-AI Bots*. Nielsen Norman Group. Retrieved January 8, 2024, from https://www.nngroup.com/articles/ai-bot-comparison/

Skjuve, M., Følstad, A., & Brandtzaeg, P. B. (2023). The User Experience of ChatGPT: Findings from a Questionnaire Study of Early Users. *Proceedings of the 5th International Conference on Conversational User Interfaces*, 1–10. https://doi.org/10.1145/3571884.3597144

Tankala, S. (2022, November 27). Why and How to Use Demographics in UX. Nielsen Norman Group. https://www.nngroup.com/articles/demographics-in-ux/

Tong, A. (2023, September 7). *Exclusive: ChatGPT traffic slips again for the third month in a row.* Reuters. Retrieved January 8, 2024, from https://www.reuters.com/technology/chatgpt-trafficslips-again-third-month-row-2023-09-07/

Xu, R., Feng, Y., & Chen, H. (2023). ChatGPT vs. Google: A Comparative Study of Search Performance and User Experience. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4498671

Appendix

[1] Pre-Test Questionnaire Link

[2] Pre-Test Questionnaire Results

Participant ID	Gender	Age	Education Level	Occupation	Nationality	Residency
1	Male	20 - 29	Master's Degree	Student	Syria	Turkey
2	Female	20 - 29	Master's Degree	Computer Engineer	Turkey	Turkey
3	Male	20 - 29	Bachelor's Degree	Intern	Turkey	Turkey
4	Male	20 - 29	Bachelor's Degree	Engineer	Turkey	Turkey
5	Female	20 - 29	Master's Degree	Product Manager / Student	Turkey	Turkey
6	Male	20 - 29	Master's Degree	Financial Analyst	Turkey	Turkey
7	Female	20 - 29	Bachelor's Degree	Student	Turkey	Turkey

8	Male	20 - 29	Master's Degree	Student	Turkey	Turkey
9	Male	20 - 29	Bachelor's Degree	Data Scientist	Turkey	Turkey
10	Female	20 - 29	Bachelor's Degree	Student	Turkey	Turkey
11	Female	30 - 39	Doctorate Degree	Researcher	Turkey	Turkey

Participant ID	**Overall Computer Skills (1-5 Scale)**	**Previous experience with AIBased Search Engines**	**Opinion About AI-Based Search Engines in General (1-5 Scale)**	**Interest in AI-Based Search Engines (1-5 Scale)**	**Perceived Visual Ability (1-5 Scale)**	**Physical or Cognitive Impairments**
1	4	Yes	5	5	2	No
2	5	No	5	5	4	No
3	4	Yes	4	4	4	No
4	5	No	4	5	5	No
5	5	Yes	5	3	5	No
6	3	No	3	3	5	No
7	5	Yes	5	5	5	No
8	2	No	4	2	5	No
9	3	Yes	4	5	4	No
10	5	Yes	5	4	3	No
11	3	No	4	4	4	No

[3] SUS Link

[4] Post-Test Questionnaire Link

[5] OLS Regression Model Results for Bing

	Coef	**Std-Err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Const	15,06	5,48	2,75	0,11	-8,50	38,61
Age	-0,40	0,23	-1,75	0,22	-1,38	0,58
Interest	-0,42	0,24	-1,71	0,23	-1,47	0,63
Trust	-0,12	0,25	-0,49	0,67	-1,18	0,94
Immersion	-0,35	0,20	-1,77	0,22	-1,21	0,50
Gender (Male)	1,68	0,39	4,35	**0,044****	0,02	3,34
Education(Doctorate)	2,14	1,08	1,97	0,19	-2,53	6,80

Education (Masters)	0,04	0,59	0,07	0,95	-2,51	2,59
Occupation(Researcher)	2,14	1,08	1,97	0,19	-2,53	6,80
Occupation(Student)	-0,02	0,31	-0,08	0,95	-1,35	1,31