Investigating the User Experience and Evaluating Usability Issues in AI-Enabled Learning Mobile Apps: An Analysis of User Reviews

Bassam Alsanousi[1], Abdulmohsen S. Albesher[2], Hyunsook Do[3], Stephanie Ludi[4]
Computer Science and Engineering, University of North Texas, Denton, TX, USA[1,3,4]
Information Systems, King Faisal University, Alahsa, Saudi Arabia[2]

***Abstract*—Integrating artificial intelligence (AI) has become crucial in modern mobile application development. However, the current integration of AI in mobile learning applications presents several challenges regarding mobile app usability. This study aims to identify critical usability issues of AI-enabled mobile learning apps by analyzing user reviews. We conducted a qualitative and content analysis of user reviews for two groups of AI apps from the education category - language learning apps and educational support apps. Our findings reveal that while users generally report positive experiences, several AI-related usability issues impact user satisfaction, effectiveness, and efficiency. These challenges include AI-related functionality issues, performance, bias, explanation, and ineffective Features. To enhance user experience and learning outcomes, developers must improve AI technology and adapt learning methodologies to meet users' diverse demands and preferences while addressing these issues. By overcoming these challenges, AI-powered mobile learning apps can continue to evolve and provide users with engaging and personalized learning experiences.**

***Keywords*—Human-Computer Interaction (HCI); Artificial Intelligence (AI); user reviews; AI-Enabled Mobile Apps; usability**

I. Introduction

There is a growing demand for mobile apps incorporating artificial intelligence (AI) as more people use them to enhance their daily lives. The size of the AI apps market is expected to expand rapidly in the coming years [1]. By incorporating AI into mobile apps, programmers can introduce cognitive and logical characteristics that lead to a diverse selection of smartphone applications. [2]. Implementing AI technology is prevalent across various facets of human existence, particularly within the field of education [3].

Mobile learning applications incorporating AI provide a sophisticated educational atmosphere, progressive pedagogical techniques, and easily accessible platforms for learners [4]. Recent studies have noted the positive impact of AI-based in English education [5]. The increasing trend of mobile assisting language learning apps has resulted in the utilization of AI-powered speaking applications equipped with speech evaluation mechanisms to facilitate English as a foreign language (EFL) speaking exercises [6]. Besides, the current wave of artificial intelligence is already impacting the management and domination of math education apps [7].

However, AI systems must demonstrate effectiveness and efficiency while ensuring user satisfaction. This is crucial because usability strongly influences the success and quality of software [8]. Therefore, to evaluate the usability issues in AI-enabled learning mobile apps, this research will utilize usability key factors such as user satisfaction, efficiency, and effectiveness, commonly considered when assessing mobile apps [9] [10].

The rapid expansion of mobile devices has resulted in an upsurge in mobile applications accompanied by user reviews [11]. In mobile app marketplaces like the Apple Store and Google Play, users can rate apps and post reviews about adding a feature, reporting an issue, and their experiences using them. These reviews can be rich sources of information for developing future software versions [12]. In other words, developers can improve software quality and identify missing features by analyzing user reviews [11]. The previous study has shown that these reviews possess information related to user usability [13]. Thus, analyzing app reviews is a helpful way to identify any usability problems that users might face and to reveal any looked-for improvements [14]. Groen et al. [15] discovered that these app reviews have the ability to discover quality features directly affecting users.

Despite the wide adoption of AI-enabled learning mobile apps, there is still a need to better understand the specific usability issues associated with these apps. The purpose of this study is to address this gap. Additionally, while several studies [11] [14] [16] [17] [18] analyzed user reviews to detect usability issues to enhance software quality, there is insufficient research regarding the specific examination of AI-enabled learning apps. This is an area that our study aims to investigate and address.

In response to the identified gaps, this research makes notable contributions to the field of AI-enabled mobile learning apps. Through a comprehensive analysis of user reviews focusing on user experience and usability issues, our study offers a robust understanding that can assist app developers and designers in improving the design and development of AI-enabled mobile learning apps. Understanding user perspectives can address usability issues, improve user satisfaction, increase effectiveness, and enhance efficiency in mobile learning experiences. The research findings also contribute to the broader understanding of AI in education, highlighting the challenges and opportunities associated with AI integration in mobile learning apps, thus facilitating their wider acceptance and use among learners and educators.

This study aims to perform a thorough usability evaluation

of AI-enabled mobile learning apps, following ISO 9241 standards [9] [10], measuring effectiveness, efficiency, and satisfaction. The primary objective is to identify potential usability issues affecting user satisfaction, effectiveness, and efficiency. Additionally, to identify the primary challenges and deficiencies that AI-enabled learning applications encounter in delivering satisfactory performance. This study provides valuable information for application developers and designers to improve app usability by analyzing user reviews of various AI-enabled mobile applications in the education category. The findings help identify the apps with significant problems and weaknesses, showing where improvements should be focused. Based on these goals, this research aims to answer the following research questions:

RQ1: How is the user experience with AI-enabled mobile learning apps?

RQ2: To what extent do usability issues impact user satisfaction, effectiveness, and efficiency in AI-enabled mobile learning apps?

RQ3: What are the most prevalent usability issues related to AI in AI-enabled mobile learning apps?

The responses to the research questions mentioned above will offer crucial insights into the most prevalent usability issues related to AI-enabled mobile applications for language learning and educational support. These insights can be leveraged to identify areas that require improvement and formulate effective strategies to enhance the usability of these apps. Ultimately, this research aims to improve the usability of AI-enabled mobile learning apps, thereby enhancing user satisfaction, effectiveness, and efficiency.

The paper is organized as follows: Section II presents usability evaluation in mobile applications, and leverages user reviews for detecting usability issues in diverse mobile applications. Then, Section III explains and justifies the research methodology utilized in this study. Section IV presents the results, while Section V engages in the discussion of the results. Section VI focuses on addressing potential threats to validity. Finally, Section VII concludes the paper and discusses future work.

II. Related Work

This section highlights the most usability evaluation in mobile apps and the utilization of user reviews to detect usability problems in mobile apps.

A. Usability Evaluation in Mobile Application

Many studies have conducted usability evaluations of mobile apps [15] [19] [20]. For instance, a systematic review by [19] showed a literature review on a comprehensive examination of the usability of mobile apps. They found that the definition of ISO 9241-11 has been mostly used in HCI, followed by ISO 25010 definition. Furthermore, they discovered that efficiency, satisfaction, and effectiveness were the most frequent attributes. Another systematic study by Sunardi et al. [20] introduced a literature review of the most usability evaluations. They observed that satisfaction, efficiency, and effectiveness are the most well-known usability evaluations. Another systematic review by [21] discussed mobile app usability in different mobile app categories. They identified frequent usability for mobile applications and required design features for a particular mobile application category. Another study by [22] proposed a model that aims to employ opinion mining to evaluate the subjective usability of the software automatically. This model is centered on three crucial quality factors of usability: effectiveness, efficiency, and satisfaction. Moreover, Alhejji et al. [23] completed a comprehensive analysis, assessment, and comparison of the usability of mobile banking apps, considering both iOS and Android platforms in Saudi Arabia. The researchers evaluated usability based on effectiveness, efficiency, and satisfaction, as defined by ISO 9241.

B. Leveraging user Reviews for Detecting Usability Issues in Diverse Mobile Applications

Previous studies used user reviews to extract valuable information [11] [14] [16] [17]. Felwah and Rita [14] analyzed 1236 reviews from 106 mental health apps to detect usability issues. They categorized usability issues into six groups, finding that the results could offer app developers valuable design advice to enhance the usability of mental health apps. The further study analyzed [16] user reviews in stroke caregiving applications to identify usability problems. They found some usability issues such as errors, efficiency, and effectiveness because of the misunderstanding of the user needs of app developers. The authors categorized the user reviews as negative or neutral using the usability evaluation criteria of Nielsen and Bevan, as well as considering the extent of the user experience. In addition, another work by [17] reviewed mobile applications created for the COVID-19 pandemic. The writers gathered feedback from users in the form of ratings and reviews as well as information on the objectives and features of the app. The review's results emphasized the need for new application development and further improvement, revealing design features like ease to use, cultural sensitivity, usefulness, privacy, responsiveness, security, flexibility, support, performance, and reliability. Another work by Pawel and Anna [11] investigated the potential of user reviews to extract usability and user experience problems in WhatsApp. They identified seven usability factors that were connected to the issues that were reported. Using different dictionary sources, they used a sentiment analysis by grouping all the negative and positive words from the user reviews. Another research study by [24] proposed employing machine learning (ML) techniques to ascertain users' perspectives. Furthermore, they utilized and compared the effectiveness of five different ML classifier algorithms. Lastly, [18] explored user reviews and discovered novel usability issues associated with disaster applications by adapting a pre-existing usability framework. The existing research on usability in mobile app user reviews has been extensive. However, the usability of AI-enabled mobile applications for language learning and educational support has received limited attention in previous research. Therefore, this study aims to analyze user reviews and uncover potential usability issues associated with AI mobile applications in this specific domain.

III. Methodology

Our motivation is to empirically evaluate usability issues in AI-enabled mobile applications designed for language learning

TABLE I. APP'S STATISTICS

App Group	App Name	Number of Reviews
Language Learning	ELSA: Learn And Speak English	45,576
	Duolingo: language lessons	276,819
	Cake - Learn English & Korean	80,309
Educational Support	Socratic by Google	5,789
	Microsoft Math Solver	3,651
	Symbolab: Math Problem Solver	1,248
	Photomath	22,1559

and educational support. To accomplish this, we have devised an approach comprising six steps, which we will refer to as A, B, C, D, E, and F, as shown in Fig. 1. This section provides a detailed description of each step.

A. Identifying AI Mobile Apps Designed for Self-Education

Our objective is to assess the usability issues in mobile applications designed for language learning and educational support using AI. To achieve this, we have identified a specific subset selection criteria to filter out the relevant apps. These selection criteria are outlined below.

- **Initial Selected Apps**: We selected the top 100 worldwide downloaded mobile applications in the education category. We sourced this information from the ranking lists on data.ai[1]. This selection aimed to include the widely used apps in our study.
- **AI-Enabled Mobile Apps:** Then, we manually investigated the description of each initialed selected top 100 apps in the education category to ensure whether an app uses AI features, as shown in Fig. 2.
- **Number of App Reviews:** We excluded apps with reviews less than 1000.

This study identified mobile applications in the education category that explicitly incorporated AI technology. we aimed to investigate the usability issues associated with apps that used AI for educational purposes. After reviewing the top 100 apps in the education category based on the selection criteria, we found that only thirteen apps met our selection criteria. Since most of our study heavily relies on user reviews, we want to ensure these apps have sufficient user reviews. Therefore, we discarded apps with reviews of less than 1000. Additionally, we excluded apps unrelated to language learning and educational support to ensure our analysis focused on relevant apps. As a result, this allowed us to filter out seven AI-enabled mobile applications, namely: *ELSA: Learn And Speak English, Duolingo: language lessons, Cake - Learn English & Korean, Socratic by Google, Microsoft Math Solver, Symbolab: Math Problem Solver and Photomath*. Although these apps belong to the same category, they serve different purposes. Consequently, we classified them into two groups - language learning and educational support, as shown in the Table I.

B. Data Collection and Cleaning

After identifying seven AI mobile applications for learning languages and educational support, we collected all available app reviews for these apps from Google Play Store. To do so, we utilized *Google-Play-Scraper*[2] using Python to crawl user reviews for each app. We collected **634,951** user reviews for all seven apps between March 2022 and March 2023. However, we found that the initial data contained irrelevant and noisy information, such as short sentences, emojis, and non-English reviews, which could be problematic when answering our research questions. Therefore, we used Natural Language Processing (NLP) libraries to make our results more reliable by removing all irrelevant and noisy data. We utilized the following cleaning criteria to filter out irrelevant reviews:

1) Blank content.
2) Emojis-only content.
3) Emojis and numbers from the content.
4) Duplication content.
5) Non-English.
6) Less than two words.

As a result, we exclude 189,491 irrelevant app reviews with the remaining **445,460** relevant reviews.

C. Sentiment Analysis

We conducted sentiment analysis to filter out reviews to positive, negative, and neutral as it offers advantages in identifying usability [25]. To determine the sentiment of each review sentence, we utilized a domain-specific sentiment analysis tool, SentiStrength-SE [26], specifically designed for analyzing text in software engineering, including analyzing user reviews [27]. The SentiStrength-SE algorithm evaluates each word individually and assigns a score to indicate its sentiment. The tool allocated numerical values to positive and negative words, ranging from +1 to +5 for positive and -1 to -5 for negative [28]. The polarity of sentiment, as determined by its value, can be categorized as negative when the value is less than 0, positive when the value is greater than 0, or neutral when the value is equal to 0. This classification allows for assessing the severity of the sentiment based on its proximity to 0. This approach, as described in [29], provides a framework for understanding the sentiment expressed in user reviews. To compute the overall sentiment score for each sentence, we determined the highest positive and negative scores and combined them [28]. By utilizing the SentiStrength-SE tool on **445,460** reviews, it filtered them based on three categories (Neutral = 163,142, Positive = 247,640, and **Negative = 34,678**). This process is beneficial for collecting candidate reviews for our evaluation in the next step and reduces the number of reviews that will be matched with usability factors: satisfaction, effectiveness, and efficiency [22].

D. Usability Keywords Filtering

The purpose of using usability keyword filtering in this step is to identify potential usability issues quantitatively by analyzing user reviews. However, the reviews we obtained contained mixed feedback, including positive, neutral, and negative opinions. We only analyzed the negative reviews since

[1]https://www.data.ai/

[2]https://pypi.org/project/google-play-scraper/

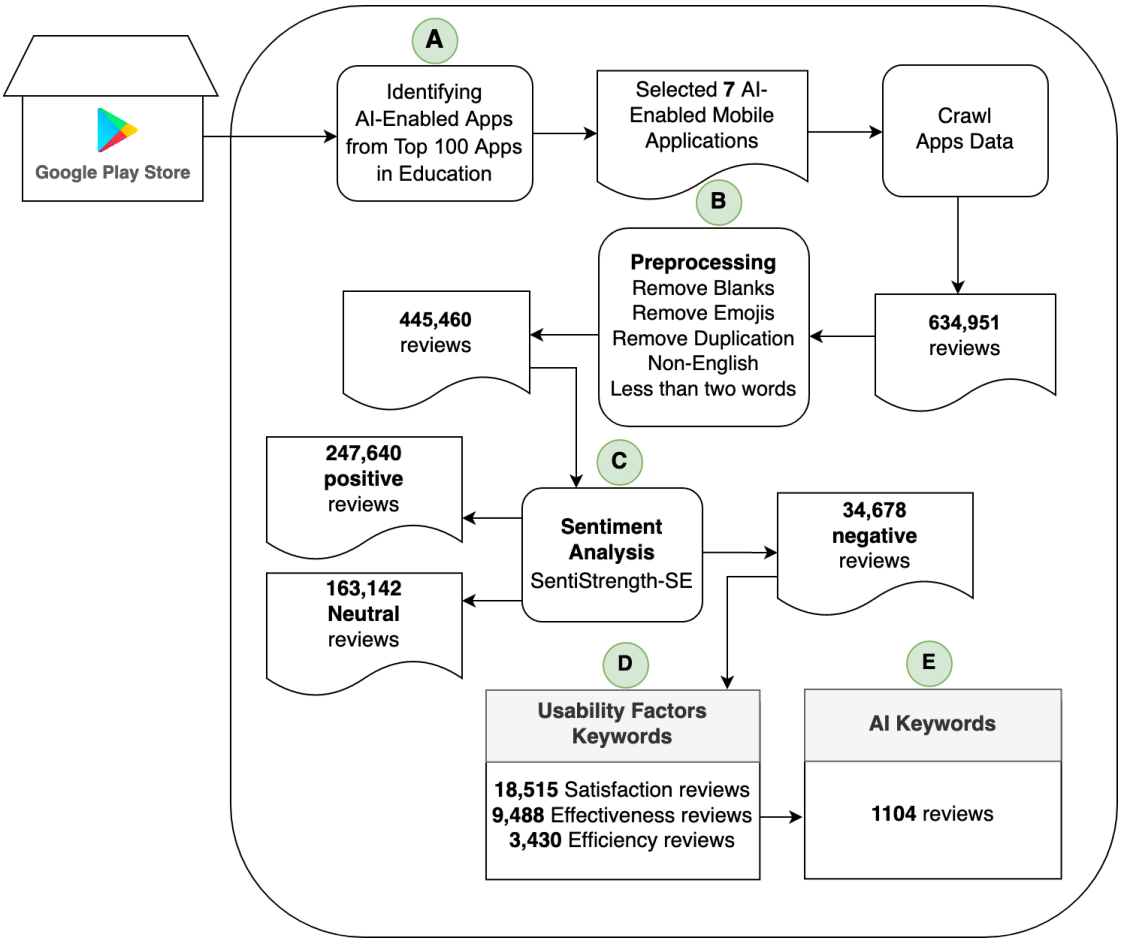


Fig. 1. Overview approach.

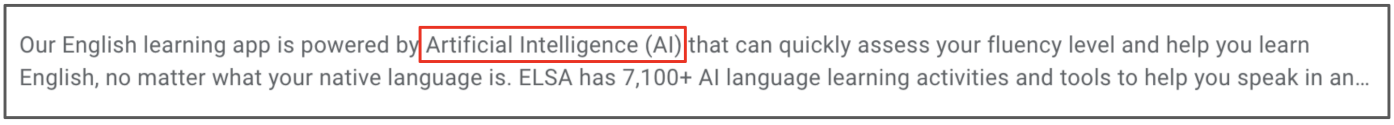


Fig. 2. App's description.

we focused on examining the usability issues in AI-enabled mobile learning. As defined by ISO 9241-11, Usability refers to how effective and efficient a product is and the degree of user satisfaction after using it [30]. To better understand the usability issues associated with these apps, we constructed three negative polarity lexicons based on satisfaction, effectiveness, and efficiency keywords. These keywords were identified by conducting a thorough examination of relevant studies [23] [11], as shown in Fig. 3. We labeled the app reviews using a Python script that employs the NLTk library, which considers lemmatization and stemming. For example, the word "crashed" can be lemmatized to "crash," and "disappointing" can be stemmed to "disappoint". Using these techniques, we could identify and classify specific usability issues more effectively, leading to more accurate results. Next, we calculated the count and frequency of negative reviews of each usability factor [9]. This resulted in 18,515 reviews related to satisfaction, 9,488 reviews on effectiveness, and 3,430 on efficiency. We then determined the overall usability score for each app review by adding the negative satisfaction, effectiveness, and efficiency scores together [10] [30]. We evaluated the usability issues of AI-enabled mobile apps for language learning and educational support by comparing the usability scores. Then, we identified which apps had the most usability issues and which had minor ones.

E. AI Keywords Filtering

After identifying usability issues from the previous step, we combined the results of 28,948 relevant negative usability reviews to filter out reviews based on AI terms [31], as shown in Fig. 4, to find only AI-related reviews. To achieve this, we developed another Python script that employs the NLTk library, which considers lemmatization and stemming, as before. This allows for more effective identification and classification of specific AI-related usability issues, ensuring more accurate results. For example, "speech recognition" can be lemmatized

to "speech recognize", and the word "voice recognition" can be stemmed to "voic recognit" The filtered result was 1104 AI-related reviews out of 28,948. We then randomly selected AI-related reviews based on their rating [25], resulting in 221 reviews. Then, quantitative data analysis was used to examine the most prevalent usability issues related to AI in AI-enabled mobile apps for language learning and educational support. The qualitative content analysis approach enabled identifying and examining how these usability issues are reflected in user experiences. We performed a thematic analysis of the data using Excel software. Based on the steps outlined in Fig. 5, we conducted a manual analysis of the sample AI-related reviews [32]:

1. Randomly select 221 AI reviews based on the rating criteria.
2. Conduct a comprehensive reading of all 221 selected reviews to gain a deep understanding of the data.
3. Identify and note any patterns or ideas that emerge from the reviews.
4. Generate initial codes from the patterns and ideas identified in the reviews.
5. Review the codes to identify overarching themes and sub-themes that capture the essence of the data.
6. Refine the results by reviewing and comparing the themes to the original data to ensure they accurately reflect the content and context of the reviews.
7. Define and label each theme and sub-theme to make it clear and understandable.
8. Provide quotes from the reviews to illustrate each theme and sub-themes.

We coded the sample reviews manually after reading them multiple times to familiarize themselves with the information. We assigned codes to significant phrases and sentences relevant to AI-related issues [32]. Additionally, as the research progressed, we modified the codes to represent the substance and context of the information accurately. Then, we analyzed the codes to identify recurring themes and sub-themes [33]. We appropriately labeled the data to reflect the content and context of the information. To uphold the accuracy and reliability of the analysis, we conducted various tests and inspections, including reviewing the coding strategy and checking inter-coder reliability with a second researcher. We addressed any inconsistencies or misunderstandings through discussion and agreement. We systematically conducted the manual analysis process, ensuring the findings' validity and reliability. The analysis followed the guidelines provided by [34].

IV. Results

Our evaluation examined the impact of usability issues on using AI-enabled mobile learning apps and their impact on user satisfaction, effectiveness, and efficiency. The usability issues are crucial to explore, as they can influence user behavior and affect the overall performance of educational apps. By analyzing user reviews of various AI-enabled mobile apps for learning languages and educational support, this study aims to illuminate the possible challenges or concerns that users encounter while using these apps and assess how these issues affect the user experience.

RQ1: How is the user experience with AI-enabled mobile learning apps?

Usability Keywords

Satisfaction:
hate, bad, worst, sad, disappointed, weak, horrible, difficult, ...

Effectiveness:
wrong, trouble, error, crash, problem, bug, glitch, ...

Efficiency:
freeze, slow, delay, late, laggy, hanging, stuck, take time ...

Fig. 3. Negative usability keywords.

AI Keywords

`artificial intelligence, machine learning, ai, deep learning, ml ...`

Fig. 4. Keywords related to AI.

RQ1 Rational: This question aims to determine the overall sentiment of user reviews towards AI-enabled mobile learning apps. The sentiment analysis will provide insight into how users experience and perceive these types of apps, helping identify potential improvement areas. Understanding the sentiment of user reviews can assist developers and designers in creating better AI-enabled mobile learning apps that align with users' needs and expectations.

RQ1 Results: As described in the methodology section, we conducted sentiment analysis to identify the sentiment expressed in user reviews. The sentiment analysis results for AI-enabled learning apps shown in Fig. 6 revealed that in the learning language app group Duolingo: language lessons had the highest number of reviews, totaling 213,307. However, the app received the highest negative reviews, with 21,441 users expressing dissatisfaction with its features. In addition, the app had 67,043 neutral reviews. ELSA Learn English, Get Fluent received 25,410 reviews, with 16,506 positive and only 797 negatives, resulting in a high percentage of positive reviews. The app also had 8,107 neutral reviews. Similarly, Cake - Learn English & Korean received 48,435 reviews, with 32,461 positive and 1,048 negative reviews, resulting in the highest positive review count in this category. The app also had 14,926 neutral reviews. The app also had 14,926 neutral reviews. Among the educational support app group, Socratic by Google received 3,744 reviews, with 1,940 being the highest positive, 217 negatives, and 1,587 natural reviews. Microsoft Math Solver and Symbolab: Math Problem Solver had fewer reviews, with 2,593 and 893 reviews, respectively. Symbolab: Math Problem Solver had the highest negative reviews, with 140 users expressing dissatisfaction with its features and 413 neutral reviews, followed by Microsoft Math Solver, with 240 negative reviews and 1147 natural reviews. Photomath

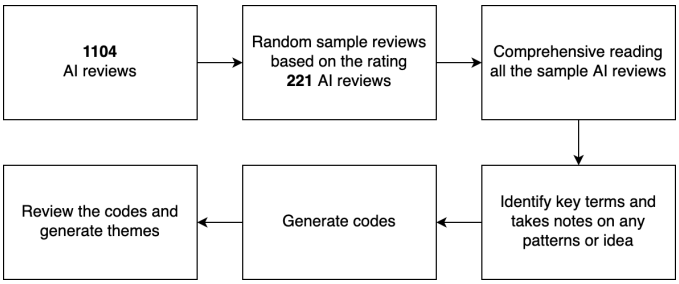
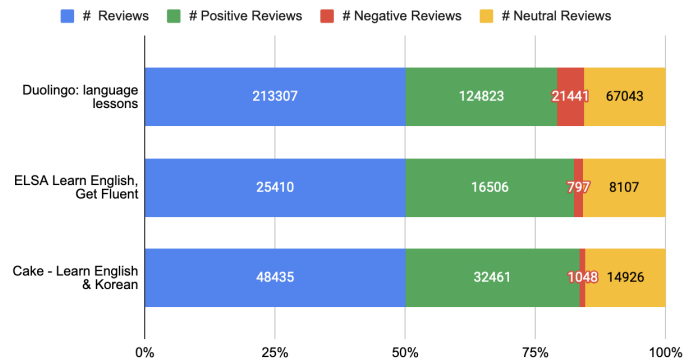
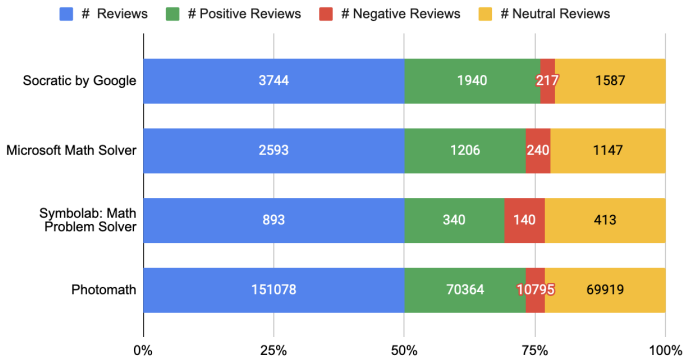


Fig. 5. A brief overview of our manual qualitative content analysis approach.



(a) Language learning apps.



(b) Educational support apps.

Fig. 6. The user experience with the seven AI-enabled mobile learning apps.

received the highest number of reviews, with 70,364 positive, 10,795 negatives, and 69,919 neutral reviews, resulting in the second-highest positive reviews count alongside Microsoft Math Solver in this category. Overall, the AI-enabled mobile learning apps had more positive than negative reviews, with 247,640 positive reviews, 34,678 negative reviews, and 163,142 neutral reviews out of 445,460 total reviews.

RQ1 Summary. This summary highlights that Duolingo had the highest number of negative reviews among language learning apps, while Cake - Learn English Korean had high percentages of positive reviews. Socratic by Google had the highest positive reviews among educational support apps, while Symbolab: Math Problem Solver had the highest negative reviews. Overall, the apps had more positive than negative reviews, with many neutral reviews.

RQ2: To what extent are the usability issues of using AI-enabled mobile learning apps impact user satisfaction, effectiveness, and efficiency?

RQ2 Rational: With the increasing prevalence of AI in mobile learning apps, it's essential to assess the impact of usability issues on user satisfaction, effectiveness, and efficiency. Investigating these factors can help identify potential problems and provide insights for developers and designers to improve these apps.

RQ2 Results: To address this question, we conducted a study that filtered out reviews based on usability factors described in the methodology section. After filtration, the word clouds of the app reviews show the most occurrences of negative usability keywords in Fig. 7. Our results in Table II indicate that usability issues impact user satisfaction, effectiveness, and efficiency in AI-enabled mobile learning apps. In the language learning apps category, ELSA Learn English and Get Fluent exhibited the highest dissatisfaction score at 61%, suggesting user discontent with the app. Duolingo: language lessons and Cake - Learn English & Korean showed quite similar dissatisfaction scores at 56% and 57%, respectively, indicating a close level of user satisfaction between the two apps. Notably, Cake - Learn English & Korean excelled with the lowest dissatisfaction scores in effectiveness and efficiency at 12% and 5%, respectively, suggesting that users find this app both effective and efficient for language learning. Both Duolingo: language lessons and ELSA Learn English and Get Fluent showed dissatisfaction scores in effectiveness at 33% and 23%, respectively. However, their efficiency dissatisfaction scores were 11% and 8%, respectively, suggesting these apps do not significantly impede users' learning efficiency. Moving to the educational support apps category, Symbolab: Math Problem Solver had the highest dissatisfaction score at 79%, indicating substantial user discontent. Microsoft Math Solver displayed a dissatisfaction score of 59% with a relatively high effectiveness dissatisfaction score of 24%. This suggests a moderate impact on user satisfaction, with a relatively lower efficiency dissatisfaction score of 6% compared to its higher effectiveness dissatisfaction score. Despite having the second highest dissatisfaction score at 63%, Socratic by Google presented low dissatisfaction scores for effectiveness and efficiency at 12% and 5%, respectively, suggesting a smaller impact on these areas. Photomath stands out with a relatively low overall dissatisfaction score of 48%, indicating better user satisfaction than other apps in this category. However, there is room for improvement as its effectiveness and efficiency dissatisfaction scores sit at 19% and 9%, respectively. Regarding the total usability score shown in Fig. 8, among the language learning apps, Duolingo: language lessons achieved the highest total usability score of 99%, indicating that users

encountered the most usability issues with this app, particularly regarding negative satisfaction and negative effectiveness. ELSA Learn English and Get Fluent obtained a total usability score of 93%, signifying significant usability issues related to satisfaction and effectiveness. On the other hand, Cake - Learn English & Korean obtained the lowest total usability score of 74%, suggesting that users encountered fewer usability issues with this app. In the educational support apps category, Symbolab: Math Problem Solver garnered the highest total usability score at 101%, indicating higher usability issues with this app compared to other apps in this category. Photomath received the lowest usability score of 75%, indicating moderate usability issues. Understanding the impact of usability issues in AI-enabled mobile learning apps can guide developers and designers in proactively identifying and resolving these problems. This process can enhance overall user satisfaction, effectiveness, and efficiency. The knowledge gained from addressing these issues will inform future improvements in the design and functionality of these apps, ultimately leading to more effective and gratifying learning experiences.

RQ2 Summary. The result showed notable usability issues with AI-enabled mobile learning apps. Among the language learning apps, Duolingo: language lessons had the highest usability issues score of 99%, while Symbolab: Math Problem Solver obtained the highest score of 101% among the educational support apps. Developers and designers can use this information to improve the overall usability of these apps.

RQ3: What are the most prevalent usability issues in AI-enabled mobile learning apps?

RQ3 Rational: This question aims to pinpoint the most frequently encountered AI-related usability problems in mobile applications designed for language learning and educational support, as reported by users in their reviews. By gaining insight into these prevalent issues, developers and researchers can concentrate on enhancing the usability aspects tied to AI technology. This knowledge will aid in refining the overall user experience and satisfaction with these apps, ultimately providing significant benefits to learners and educators.

RQ3 Results: To answer this question, we conducted a thematic analysis as described in the methodology section. We applied the thematic analysis to the two apps group learning languages apps and educational support apps. The result of the analysis is shown in Table III and outlined below:

Learning Languages Apps Group:

AI-related functionality issues: This theme refers to problems users face with the artificial intelligence components in the app. These issues can directly affect the user experience and hinder their learning process. There are two sub-themes under this main theme:

a. Voice recognition: This highlights issues where the app fails to recognize or understand the user's voice accurately, leading to frustration and a poor learning experience. A user complained that *"the app became unstable, and AI doesn't care about recognizing anything. You can sing a song instead of the correct answer, and AI will accept it. Most of the time, it freezes in the middle of the lesson, and the mic button gets stuck, among other issues. Don't buy their plan until they fix their app."* The issues related to voice recognition underline the importance of extensive testing and improvement of AI technology to ensure accuracy and effectiveness in aiding users' language acquisition journey.

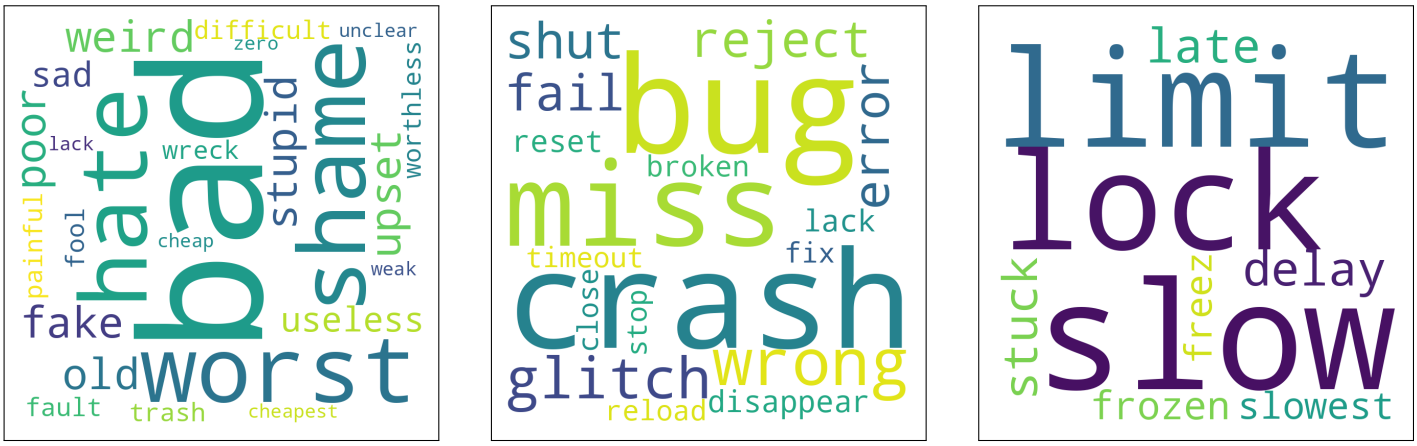
b. AI understanding of user input: This deals with situations where the app's AI fails to understand or process the user's input accurately, resulting in irrelevant or incorrect content being presented to the user, limiting the learning experience's effectiveness. One user reported, *"It was a good tool to start learning a new language from scratch. But I find that the AI component doesn't work very well. It gives easy exercises for the same word right after I've already done a more challenging one. Also, I have to practice a lot of unnecessary stuff like names and cities. A lot of practices are highly redundant, so I think I waste a lot of time practicing easy and unnecessary stuff. Furthermore, I have no choice in what to practice. All in all, it's not the most efficient."* It is crucial to integrate AI technology into language learning apps for a fulfilling educational journey.

AI Performance: This theme refers to issues related to the app's performance and capabilities of the AI component. These issues can affect the overall effectiveness of the app in helping users learn a language. There are three sub-themes under this main theme:

a. Learning methodology: This includes issues where the AI's approach to teaching a language is deemed ineffective or flawed. Users might find the exercises repetitive, redundant, or not challenging enough, which may hinder their learning progress. One user states, *"The app is basically good but its AI is too bossy. Nobody needs to be mocked for missing lessons. People get busy sometimes, and constant reminders ain't cool. They just piss people off even more."* To guarantee a satisfactory user experience, developers must design AI language learning applications with stimulating activities and lucid explanations.

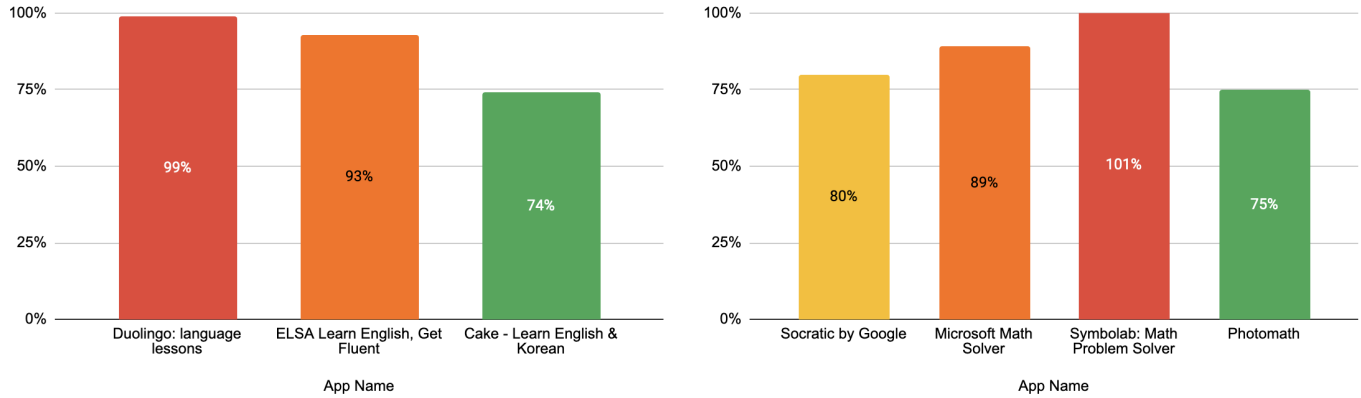
b. AI-generated content: This refers to issues with AI-generated content, such as inconsistencies, errors, or low-quality material. These problems can lead to confusion or a less engaging learning experience. One user points out, *"It's totally useless. It randomly recognizes or rejects what you're saying; there's no logic behind it. I've been using it for a while but have seen no improvement. It doesn't show you how to produce some sounds, just provides some useless text. It accepts incorrect input and randomly rejects correct input, creating an illusion of functionality. Additionally, it records all your voice data and stores it on their servers indefinitely. Even when you request deletion, the data remains."* Developers must ensure that the AI-generated content they produce is of high quality and accuracy, providing users with a positive learning experience. By focusing on improving AI-generated content, the app can better meet users' needs and enhance their language learning journey.

Instant Feedback Issues: This theme addresses problems that arise when the language learning app's AI does not offer sufficient, comprehensible guidance on the material being taught, resulting in user confusion and frustration. This main theme consists of a single sub-theme:



(a) Satisfaction (b) Effectiveness (c) Efficiency

Fig. 7. The word clouds of app reviews display the most frequent occurrences of negative usability keywords.



(a) Learning Languages Apps. (b) Educational Support Apps.

Fig. 8. The overall usability issues in the seven AI-Enabled mobile learning apps.

a. Diverse challenges with AI-generated feedback: This sub-theme pertains to situations where the app's generated content is unclear, confusing, or not provided, leading to an unsatisfactory learning experience. Without feedback, users might find it difficult to identify their errors and understand how to improve, which could impede their progress. For example, a user mentioned, *"I am learning Ukrainian from a professional teacher along with using the app Many times I am taught something that contradicts what Duolingo says Very confusing. The two voices that read out the questions also mispronounce the words I know this because I have Google translate as well as actual Ukrainians telling me that its wrong My main issue is that it does not explain why the answer is what it is If I knew the rule maybe Id get the right answer Im just frustrated and needed to vent."* It shows inaccurate or inconsistent feedback on users' pronunciation, leading to frustration and impeding their language learning progress. Addressing these concerns is crucial in enhancing the user experience and facilitating effective language learning.

Educational Support Apps Group:

AI-related Functionality Issues: This theme's main focus is on the difficulties users encounter when using the artificial intelligence elements included in the application. These challenges could reduce user satisfaction and prevent them from advancing academically. Resolving these issues could improve the efficiency and effectiveness of AI-based learning applications. The primary theme is divided into two subthemes:

a. Inaccurate problem recognition: Users report that the app fails to recognize problems, showing a different problem or failing to recognize specific symbols accurately. One user states, *"Horrible All it can do is solve basic questions and I ended up getting in a row idk how ppl like this app but it acts like ai I take a picture of the question and then it shows a totally different problem I expect you to fix this monstrosity."*

b. Poor image recognition: Users report issues with the app's image recognition capabilities, making it difficult to capture and solve problems accurately. A user shares, *"This app is complete literal trash It cant even recognize an extremely simple variable equation system properly much less solve it Leave the image recognition to the big boys like*

TABLE II. THE ANALYSIS OF OCCURRENCE AND FREQUENCY OF NEGATIVE REVIEWS FOR EACH USABILITY FACTOR

App Name	App Group	Negative Reviews	Satisfaction (%)	Effectiveness (%)	Efficiency (%)	Total Usability (%)
Duolingo: language lessons		21441	56	33	11	99
ELSA Learn English, Get Fluent	Language Learning	797	61	23	8	93
Cake - Learn English & Korean		1048	57	12	5	74
Socratic by Google		217	63	12	5	80
Microsoft Math Solver	Educational Support	240	59	24	6	89
Symbolab: Math Problem Solver		140	79	15	8	101
Photomath		10795	48	19	9	75

Google and Microsoft, You're an embarrassment to society and to yourselves."

c. Inability to solve complex problems: Users mention that the app is limited in solving complex mathematical problems or specific algorithms. As one user mentions, *"As long as this app does not solve the Gau Algorithm, it is worthless for me. Can't read matrix integrals or equations with more than one variable. Total trash."*

AI Performance: This theme draws attention to concerns with the performance and capabilities of the app's AI component, such as flaws or inconsistencies that could reduce the app's overall usefulness in aiding users in learning or problem-solving. Improved user satisfaction and learning outcomes can result from higher AI performance. This overarching theme has two sub-themes:

a. Inconsistent handwriting recognition: Users find that the app struggles to recognize handwritten problems, leading to incorrect solutions accurately. One user shares, *"Has a bit of trouble with handwriting recognition and larger, more complex problems, but those problems are to be expected from any algorithm."*

b. Limited language support: Users report that the app cannot recognize or solve problems written in certain languages, limiting its usefulness for non-English speakers. A user complains, *"Time-wasting app for Nepalese. Because it doesn't recognize the math problems properly which is written in Nepali language, so don't waste your valuable time, guys, in this app."*

Ineffective Features: This theme highlights aspects of the app that are inadequate or ineffective in helping users achieve their educational goals. Developers can enhance their app's functionality by addressing these concerns and meeting user demands.

a. Limited problem-solving capabilities: Users find the app's ability to solve certain problems insufficient or lacking. One user complains, *"This app is complete literal trash. It can't even recognize an extremely simple variable equation system properly, much less solve it."*

Instant Feedback Issues:

This refers to situations where the AI-generated feedback is either delayed, unclear, or not provided, affecting the user experience and learning process. Addressing these issues is crucial for enhancing the user experience and promoting effective learning through the app. Improving the clarity and simplicity of AI-generated content can enhance the general user experience and the effectiveness of educational support apps.

RQ3 Summary. We identified several usability issues in AI-enabled mobile language learning and educational support apps. These issues include functionality problems related to AI, performance issues, ineffective features, and lack of instant feedback. Addressing these issues is essential to improve the effectiveness of the apps and enhance the overall user experience.

a. Varied concerns with AI-generated feedback: This underline specific example where the explanations or information provided by the AI is unclear, confusing, difficult to understand, or not presented at all. One user complained, *"I am not happy with the new version the old version used to show each explanation in cost but now The AI Tutorial which was free in old version And now I have to buy in Photo math plus that saying is correct that old is gold old version is the worst now I wanted to solve one equation and wanted the AI Tutorial but now it says pay monthly or Anually education should be free I want the old version back."* Addressing these concerns is crucial to improve the user experience and facilitating learning through these apps. Focusing on the clarity and simplicity of AI-generated feedback can help relieve these concerns, eventually enhancing user experience and the effectiveness of educational support apps.

Overall, during the analysis, we identified the most prevalent usability issues related to AI in AI-enabled mobile learning applications. These issues include voice and image recognition, AI understanding of user input, poor AI performance, instant feedback issues, ineffective features, and lack of clarity in AI-generated content. It is crucial to address these issues to enhance the usability of these apps, improve user experiences, and ultimately support the success of both learners and educators.

V. RESULTS DISCUSSION

Our evaluation of user reviews indicates that AI-enabled mobile learning apps generally provide a positive user experience, but several usability issues can affect user satisfaction, effectiveness, and efficiency. Below are some key takeaways from our analysis:

Takeaway 1: Positive user experiences. Most user reviews for the apps we analyzed were positive, with users praising them for their fun, easy-to-use interface, and short, engaging lessons. For example, one user of the Cake - Learn English Korean user wrote, "Love this app! It's been helping me a lot with learning Korean. It's fun, easy, and the lessons are short,

TABLE III. THEMATIC ANALYSIS RESULTS: AI-RELATED ISSUES IN AI-ENABLED MOBILE LEARNING APPS

App Group	Theme	Sub theme
Language Learning	AI-related functionality issues	a. Voice recognition b. AI understanding of user input
	AI Performance	a. Learning methodology b. AI-generated content
	Instant Feedback Issues	a. Varied concerns with Al-generated feedback
Educational Support Apps	AI-related functionality issues	a. Inaccurate problem recognition b. Poor image recognition c. Inability to solve complex problems
	AI Performance	a. Inconsistent handwriting recognition b. Limited language support
	Ineffective Features	a. Limited problem-solving capabilities
	Instant Feedback Issues	a. Varied concerns with Al-generated feedback

which is perfect for my busy schedule." Another user of the Photomath app noted, "Really excellent mind-blowing capture information calculation step-by-step procedure brilliant app. Everyone must need this app." These positive reviews suggest that AI-enabled mobile learning apps can be an effective and enjoyable way for users to learn new languages or receive educational support.

Takeaway 2: AI-related usability issues. Our analysis revealed several AI-related usability issues that can negatively impact user satisfaction, effectiveness, and efficiency. Some of the issues are voice recognition from learning apps and image recognition from educational support apps. For example, one user of the Duolingo app noted, "The voice recognition sometimes makes mistakes, which can be frustrating." Addressing these issues will be critical for improving the future usability of AI-enabled mobile learning apps. Another example from one of the users of the Photomath app noted, "This app is complete literal trash It cant even recognize an extremely simple variable equation system properly much less solve it Leave the image recognition to the big boys like Google and Microsoft Youre an embarrassment to society and to yourselves." Addressing these issues will be critical for improving the future usability of AI-enabled mobile learning apps.

Takeaway 3: Diverse challenges in AI performance and accuracy. Our analysis revealed that users reported various AI performance and accuracy issues in AI-enabled mobile learning apps. These challenges spanned from language understanding and voice recognition to problem-solving capabilities, image recognition, and inconsistent handwriting recognition. For instance, a user of the Photomath app observed, "Inaccurate problem recognition, poor image recognition, and inability to solve complex problems." Moreover, there were issues with limited language support and speech recognition. These problems can adversely affect the learning experience and the effectiveness of the applications. It underscores the critical need for continuous development and improvement in AI technology to address these performance issues, thereby enhancing the user experience.

Takeaway 4: Provide AI explanation. With the advancement of AI, previous studies have developed tools and libraries that aim to explain the behavior and output of AI models [35]. This feature is essential, particularly from the user's perspective. Our analysis shows that unexplained AI predictions contribute to users' frustration and make it difficult for them to understand such predictions. For example, one user of the Socratic by Google app noted, "I wish there was a way to see how the AI is coming up with its solutions. Sometimes it's not clear why it's giving me the answer it is." This user's comment illustrates how users can become frustrated when the AI model does not explain its behavior when recognizing or rejecting user input. By providing AI explanation frameworks, developers can provide the AI output to the users and explain how and why the AI decided to make that output, which can increase user satisfaction.

Takeaway 5: Reduce Bias in AI. Even though the root cause of bias comes from the data itself [36], since data engineering is an essential step in machine learning, it also affects the model functionality and the degree of biased output. Addressing these biases is essential for improving AI-enabled mobile learning apps. For example, one user of the Duolingo app noted, "The app's AI tends to favor certain accents and pronunciations over others, making it difficult for learners with different accents to get accurate feedback." Another user of the ELSA app noted, "Please add british accent AI I want it so bad." These examples highlight how biases in AI can negatively impact users' experiences and learning outcomes. By addressing these biases, developers can create more inclusive and effective AI-enabled mobile learning apps.

Takeaway 6: Addressing ineffective AI features. Our analysis also identified several AI features that were deemed ineffective or flawed, such as limited problem-solving capabilities, limited language support, and repetitive or redundant exercises. For example, a user voiced dissatisfaction "The app's problem-solving AI is quite limited and often fails to provide accurate solutions. It struggles with complex math equations and frequently gives incorrect answers ..." To mitigate these challenges and improve the user experience, it is necessary to

enhance the application's AI technology or refine the learning methodology.

Despite the generally positive assessments of AI-enabled mobile learning apps, our investigation has revealed critical findings that highlight areas for development. To improve user satisfaction, effectiveness, and efficiency, developers must address usability challenges related to AI performance, accuracy, bias, and explanation. By solving these issues, developers can create more effective and inclusive mobile learning apps that utilize AI to provide engaging and personalized learning experiences. Moreover, enhancing AI technology and adapting learning paradigms to meet users' diverse demands and preferences is essential. These measures will ensure that AI-powered mobile learning apps continue to evolve and offer students engaging learning opportunities.

VI. Threats to Validity

Evaluating the usability of AI-enabled mobile apps for language learning and educational support involves data gathering, filtering, and manual classification, which can be susceptible to various threats that may impact the results.

Internal Validity Regarding internal threats to our evaluation, one concern is the accuracy of the thematic analysis and the coding process, particularly in matching AI reviews with appropriate themes. This process is susceptible to human error and incorrect matches. To address this issue, we employed a rigorous thematic analysis approach in which two authors independently coded the reviews and established a code of agreement. The agreement was measured using a scale of 1 for strongly agree, 0 for neutral, and -1 for disagree. We only included reviews in our analysis that both authors strongly agreed on to minimize the potential for error and ensure the reliability of our findings.

External Validity Regarding external validity, our findings may need to be more generalizable. The data collection process only included mobile apps from the Android platform. Therefore, the results of this study may not apply to other mobile platforms, such as the Apple App Store.

VII. Conclusion and Future Work

In conclusion, AI-enabled mobile learning apps have shown great potential to provide users with effective learning experiences. However, addressing usability issues related to AI-related functionality, performance, bias, explanation, and ineffective features is crucial to enhancing user satisfaction, effectiveness, and efficiency. Developers must prioritize enhancing specific AI technologies and adapting learning methodologies to cater to users' diverse needs and preferences. By implementing these improvements, AI-powered mobile learning apps can become more inclusive and effective, leading to engaging and personalized learning experiences for users and fostering a promising future for AI-enabled mobile learning applications.

References

[1] J. Gao, P. H. Patil, S. Lu, D. Cao, and C. Tao, "Model-based test modeling and automation tool for intelligent mobile apps," in *2021 IEEE International Conference on Service-Oriented System Engineering (SOSE)*. IEEE, 2021, pp. 1–10.

[2] A. Sircar, G. Tripathi, N. Bist, K. A. Shakil, and M. Sathiyanarayanan, "Emerging technologies for sustainable and smart energy," 2022.

[3] M. E. Dogan, T. Goru Dogan, and A. Bozkurt, "The use of artificial intelligence (ai) in online learning and distance education processes: A systematic review of empirical studies," *Applied Sciences*, vol. 13, no. 5, p. 3056, 2023.

[4] K. Mohiuddin, M. N. Miladi, M. Ali Khan, M. A. Khaleel, S. Ali Khan, S. Shahwar, A. Nasr, and M. Aminul Islam, "Mobile learning new trends in emerging computing paradigms: An analytical approach seeking performance efficiency," *Wireless Communications and Mobile Computing*, vol. 2022, 2022.

[5] L. Ikawati, A. Z. Rahimi, F. Khairunnisa, M. I. Fauzan, and S. Rahayu, "Efl students' perceptions on duolingo: How ai can eliminate socioeconomic discrepancies," *EDULANGUE*, vol. 5, no. 2, pp. 254–269, 2022.

[6] B. Zou, X. Guan, Y. Shao, and P. Chen, "Supporting speaking practice by social network-based interaction in artificial intelligence (ai)-assisted language learning," *Sustainability*, vol. 15, no. 4, p. 2872, 2023.

[7] J. T. Hertel, "Algorithms and mathematics education a response and review of hannah fry's hello world: Being human in the age of algorithms," *The Mathematics Enthusiast*, vol. 20, no. 1, pp. 139–151, 2023.

[8] E. Bakiu and E. Guzman, "Which feature is unusable? detecting usability and user experience issues from user reviews," in *2017 IEEE 25th international requirements engineering conference workshops (REW)*. IEEE, 2017, pp. 182–187.

[9] M. Alghareeb, A. S. Albesher, and A. Asif, "Studying users perceptions of covid-19 mobile applications in saudi arabia," *Sustainability*, vol. 15, no. 2, 2023.

[10] S. Alhejji, A. Albesher, H. Wahsheh, and A. Albarrak, "Evaluating and comparing the usability of mobile banking applications in saudi arabia," *Information*, vol. 13, no. 12, 2022.

[11] P. Weichbroth and A. Baj-Rogowska, "Do online reviews reveal mobile application usability and user experience? the case of whatsapp," in *2019 Federated Conference on Computer Science and Information Systems (FedCSIS)*. IEEE, 2019, pp. 747–754.

[12] T. Wang, P. Liang, and M. Lu, "What aspects do non-functional requirements in app user reviews describe? an exploratory and comparative study," in *2018 25th Asia-Pacific Software Engineering Conference (APSEC)*. IEEE, 2018, pp. 494–503.

[13] S. Hedegaard and J. G. Simonsen, "Extracting usability and user experience information from online user reviews," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2013, pp. 2089–2098.

[14] F. Alqahtani and R. Orji, "Usability issues in mental health applications," in *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*, 2019, pp. 343–348.

[15] E. C. Groen, S. Kopczyńska, M. P. Hauer, T. D. Krafft, and J. Doerr, "Users—the hidden software product quality experts?: A study on how app users report quality aspects in online reviews," in *2017 IEEE 25th international requirements engineering conference (RE)*. IEEE, 2017, pp. 80–89.

[16] E. H. Lobo, M. Abdelrazek, A. Frølich, L. J. Rasmussen, P. M. Livingston, S. M. S. Islam, F. Kensing, and J. Grundy, "Detecting usability and user experience issues in stroke caregiving apps: an analysis of user reviews," 2022.

[17] M. N. Islam, I. Islam, K. M. Munim, and A. N. Islam, "A review on the mobile applications developed for covid-19: an exploratory analysis," *Ieee Access*, vol. 8, pp. 145 601–145 610, 2020.

[18] M. L. Tan, R. Prasanna, K. Stock, E. E. Doyle, G. Leonard, and D. Johnston, "Modified usability framework for disaster apps: a qualitative thematic analysis of user reviews," *International Journal of Disaster Risk Science*, vol. 11, no. 5, pp. 615–629, 2020.

[19] P. Weichbroth, "Usability of mobile applications: a systematic literature study," *IEEE Access*, vol. 8, pp. 55 563–55 577, 2020.

[20] G. F. P. Desak *et al.*, "List of most usability evaluation in mobile application: A systematic literature review," in *2020 International Conference on Information Management and Technology (ICIMTech)*. IEEE, 2020, pp. 283–287.

[21] Z. Huang and M. Benyoucef, "A systematic literature review of mobile application usability: addressing the design perspective," *Universal Access in the Information Society*, pp. 1–21, 2022.

[22] A. M. El-Halees, "Software usability evaluation using opinion mining." *J. Softw.*, vol. 9, no. 2, pp. 343–349, 2014.

[23] M. Booday and A. Albesher, "Evaluating the usability of mobile applications: The case of covid-19 apps in saudi arabia," in *2021 22nd International Arab Conference on Information Technology (ACIT)*. IEEE, 2021, pp. 1–7.

[24] O. Oyebode, F. Alqahtani, and R. Orji, "Using machine learning and thematic analysis methods to evaluate mental health apps based on user reviews," *IEEE Access*, vol. 8, pp. 111 141–111 158, 2020.

[25] L. d. N. Diniz, J. C. de Souza Filho, and R. M. Carvalho, "Can user reviews indicate usability heuristic issues?" in *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, 2022, pp. 1–6.

[26] M. R. Islam and M. F. Zibran, "Leveraging automated sentiment analysis in software engineering," in *2017 IEEE/ACM 14th International Conference on Mining Software Repositories (MSR)*. IEEE, 2017, pp. 203–214.

[27] Y. Wang, J. Wang, H. Zhang, X. Ming, L. Shi, and Q. Wang, "Where is your app frustrating users?" in *Proceedings of the 44th International Conference on Software Engineering*, 2022, pp. 2427–2439.

[28] R. Kaur, K. K. Chahal, and M. Saini, "Analysis of factors influencing developers' sentiments in commit logs: Insights from ap," *Software Engineering Journal*, vol. 16, no. 1, 2022.

[29] S. F. Huq, A. Z. Sadiq, and K. Sakib, "Understanding the effect of developer sentiment on fix-inducing changes: An exploratory study on github pull requests," in *2019 26th Asia-Pacific Software Engineering Conference (APSEC)*. IEEE, 2019, pp. 514–521.

[30] M. Alghareeb, A. S. Albesher, and A. Amna, "Studying users' perceptions of COVID-19 mobile applications in saudi arabia," vol. 15, no. 2.

[31] M. Estévez Almenzar, D. Fernández Llorca, E. Gómez, and F. Martinez Plumed, "Glossary of human-centric artificial intelligence," Joint Research Centre (Seville site), Tech. Rep., 2022.

[32] M. A. Alismail and A. S. Albesher, "Evaluating developer responses to app reviews: The case of mobile banking apps in saudi arabia and the united states," *Sustainability*, vol. 15, no. 8, 2023.

[33] M. R. Haque and S. Rubya, "" for an app supposed to make its users feel better, it sure is a joke"-an analysis of user reviews of mobile mental health applications," *Proceedings of the ACM on Human-Computer Interaction*, vol. 6, no. CSCW2, pp. 1–29, 2022.

[34] V. Braun and V. Clarke, "Using thematic analysis in psychology," *Qualitative research in psychology*, vol. 3, no. 2, pp. 77–101, 2006.

[35] F. Xu, H. Uszkoreit, Y. Du, W. Fan, D. Zhao, and J. Zhu, "Explainable ai: A brief survey on history, research areas, approaches and challenges," in *CCF international conference on natural language processing and Chinese computing*. Springer, 2019, pp. 563–574.

[36] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021.