Kavy: Fostering Language Speaking Skills and Self-Confidence Through Conversational AI

Sankha Cooray
sankha@ahlab.org
Augmented Human Lab, National University of Singapore
Singapore

Chathuranga Hettiarachchi
chathuranga@cse.mrt.ac.lk
University of Moratuwa
Sri Lanka

Vishaka Nanayakkara
vishaka@cse.mrt.ac.lk
University of Moratuwa
Sri Lanka

Denys J.C. Matthies
denys.matthies@th-luebeck.de
Technical University of Applied Sciences Lübeck, Fraunhofer IMTE
Lübeck, Germany

Yasith Samaradivakara
yasith@ahlab.org
Augmented Human Lab, National University of Singapore
Singapore

Suranga Nanayakkara
suranga@ahlab.org
Augmented Human Lab, National University of Singapore
Singapore, Singapore

Figure 1: Experience the Power of Kavy's Conversational AI, Fostering Language Skills and Confidence

ABSTRACT

Cognitive augmentation is the process of enhancing one's abilities, including learning a new language. For this, we could utilize conversational chatbots. Conventional chatbots such as *Siri*, have predominantly been based on the question-and-answer model, where a communicator seeks a specific answer to accomplish a specific task. The conversational capabilities of chatbots offer great potential to promote English language learning, particularly in developing countries, such as Sri Lanka, where many young adults lack confidence in speaking English. This is due to limited exposure to conversational-style learning and a lack of opportunity to practice without social anxiety which is often rooted in the fear of making mistakes. In this paper, we developed a conversational chatbot, Kavy, as a companion to help them practice English. We investigated, in a study with 40 users, if Kavy could improve a communicator's proficiency (e.g., verbal expression, conversation length, quality of speech) and self-confidence using both poetic and non-poetic conversational styles. We found that the users were highly motivated by the poetic version, with its use resulting in a significant increase in vocabulary. Nevertheless, a poetic chatbot may present challenges, with several users reporting that they find the poetic version confusing. We see this pioneering work as a first and promising approach that should be continued to be investigated in the future.

CCS CONCEPTS

• **Human-centered computing**; • **Computing methodologies → Natural language processing**;

KEYWORDS

Social chatbots, Conversational AI Agents, Voice Interfaces, Artificial Intelligence, Language Studies, Self Confidence, Poetry, Cognitive Augmentation

ACM Reference Format:
Sankha Cooray, Chathuranga Hettiarachchi, Vishaka Nanayakkara, Denys J.C. Matthies, Yasith Samaradivakara, and Suranga Nanayakkara. 2024. Kavy: Fostering Language Speaking Skills and Self-Confidence Through Conversational AI. In *Proceedings of The Augmented Humans International Conference (AHs 2024).* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3652920.3652944

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

AHs 2024, April 4–6, 2024, Melbourne, VIC, Australia
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0980-7/24/04
https://doi.org/10.1145/3652920.3652944

1 INTRODUCTION

Augmenting Human Intellect was Douglas Engelbart's ultimate vision [6] sketching how computer technology should be utilized in modern society. Augmenting human intellect involves enhancing an individual's ability to tackle complex problems, comprehend information tailored to their needs, and derive solutions. With prevailing globalization, a widespread challenge is the inability of many individuals to speak a second language, such as English. However, learning a new language can be challenging and time-consuming. In general, most people may have good literacy in three components of language skills which are *reading, writing,* and *listening*. Particularly, in the speaking component, beginners lack the ability to sustain a conversation. This is because they especially lack self-confidence in their belief of having insufficient vocabulary. The problem is multifaceted, with potential reasons including the lack of confidence [9], vocabulary [4], and exposure [13].

Support could be accomplished by virtual voice companions, which are widely available and suitable for assisting humans in performing various tasks. Practicing speaking in a virtual environment with a virtual assistant could potentially improve someone's speaking skills and vocabulary. However, most existing voice assistants do not encourage users to continue the conversation, as they are designed to provide factual responses for each input query. Furthermore, current implementations of voice companions are unable to handle imperfect or incomplete speech, which hinders the ability of the user to improve their language speaking skills.

In this paper, we present the development and evaluation of a mobile and personal voice companion named Kavy that is specially designed to speak poetically with a human-like voice, which can also handle imperfect and immature user input. We utilized ChatGPT for Kavy's back-end, while the front-end is a React native mobile app that runs on various platforms. We engineered prompts that generate poetic responses while also providing factual information in response to any user inquiry. With Kavy, we investigate the following research questions: *(1) How can a voice companion support the language learning progress? (2) What is the impact of a poetic version of the companion?*

We developed an artifact, a mobile app that can either behave as a poetic voice companion or a non-poetic voice companion, depending on the internal configuration. These configurations are obtained from the back-end and the configuration for each individual is assigned and saved in our database. We ran a between-subject study dividing participants into two groups named *'non-poetic'* and *'poetic'* group. The assignment is randomly performed, resulting in both groups having an equal number of participants. Participants were instructed to complete a pre-survey as well as a post-survey at the end. For two weeks, the participants were instructed to talk with Kavy as a conversational companion. They were free to talk about any topic of their choice. All user activities were automatically logged in our database, providing us the ability to perform a broad analysis on various parameters and matrices.

In this paper, the contribution is twofold. We contribute Kavy, an app together with a back-end system and integrated mechanisms to roll out surveys as one package. This is considered an artifact contribution to the HCI research community. Further, we present an empirical contribution, including an analysis of our user study, with valuable insights into how a voice companion performs for adult learners.

2 RELATED WORK

2.1 LLM poem generation

Many studies have proven the ability and power of ChatGPT poem generation using specific prompts. Studies indicate [5] that ChatGPT is capable of memorizing and retrieving famous poems as it is trained on chunks of the web. A. Dai et al. have proposed a method that trains previous GPT-2 model to generate bespoke poems curated to match with famous poet Emily Dickinson's writing styles [3]. A similar study was conducted by Sawicki P. et al. and demonstrated the usage of GPT models to generate poems in Whitman's style [21]. A.R. Kirmani [10] has documented a possible way to utilize GPT models to generate poems especially auto-crafted ones in the science domain. In summary, prior work demonstrates the capability of GPT to generate realistic poems. In our work, we adapt GPT in a similar way, but to empower Kavy as a poet with the goal of fostering conversations for language learning purposes.

2.2 Social chatbots

There is a body of work to explore the use of Artificial Intelligence (AI) conversational chatbots to engage users in meaningful conversations [1, 23]. For instance, Kaluarachchi et al. investigated how a poetic voice companion could engage people in conversations with relevant output. Microsoft developed an AI chatbot *XiaoIce* [25] that can help with people'scommunication, affection, and social belonging. These attempts demonstrate that social chatbots can be utilized to engage users in conversations. We would like to expand on these previous works by using social chatbots as an aid to learn language learning.

2.3 Poetry and self-confidence

Numerous studies discovered that the nature of poetry correlates with human self-confidence in several ways [2]. Sara Törnqvist's study [22] discovered that someone may feel less anxious speaking English when performing spoken word poetry. Furthermore, her research indicates poems to be another factor with a great impact on the students' motivation to perform. *Poetry out loud* [20] and *Poetry out loud to thinking out loud* [19] are two methodologies discovered by Saad S. et al. Their findings indicate that reading contemporary poetry has a significant impact on improving someone's confidence, oral language skills, and reading fluency and enhances public speaking abilities. All the above work demonstrate a correlation between poetry and self-confidence. Therefore, we assume that a similar approach with a slightly new perspective – conversing with a chatbot – will also positively influence the learner's self-confidence.

3 KAVY

We designed and developed an artifact, named Kavy. A mobile app with a reduced front-end to effectively focus on the conversation. The app is cross-platform compatible and works best on Android and iOS smartphones.

3.1 System Design

We integrated several existing frameworks to rapidly develop a stable and friendly-looking mobile app, serving as a vehicle research artifact. The Kavy system consists of three major components which are seamlessly connected services. The front end was developed using a cross-platform framework. It communicated with a back-end which is a NodeJS server instance that corresponds to to a MongoDB database. The overall architecture is displayed in figure 2.

3.2 Mobile app

The mobile app is developed using the React-Native framework as it provides cross-platform support to build both Android and iOS apps using a single code base. We further integrated the Google Text-to-Speech API as the main audio synthesizing service and Kavy's reply voice has been fine-tuned to adapt into a natural human voice using Neural2 voice models. To detect the spoken user input, we used the platform's native voice detection frameworks. Then, ChatGPT was used as the response generation service. The user interface intentionally adhered to minimalist design principles (less is more [8], consistent design, visual clarity [12], simplicity [17]). This was done to minimize confusion and make the app hopefully straightforward to use. A single button is to be pressed and held, which is introduced early at the home screen for easy interaction (*press and hold to speak then release to send*). The interface can be seen in figure 3.

3.3 Prompt engineering

Modern large language models have shown to be smart enough to generate complex outputs, such as tailored computer code segments, once the user sets a proper prompt. OpenAI's back-end works with JSON objects, which are used to set the model and retrieve responses. Figuring out a suitable input prompt to set the responses in a desired way is meanwhile denoted as prompt engineering [24]. Multiple prompt revisions were evaluated to achieve our optimal prompt in terms of response time and response quality. During development, we observed the delay in web request-response and determined the optimal prompt that ensures a shorter waiting time. An optimal prompt should also minimize decoding errors, which we consider a measure of quality. Below prompts are used in our Kavy app and the result takes a shorter waiting time and the quality is acceptable. Table 1 shows three examples of how Kavy responds to user input. These poetic responses encompass a diverse range of styles, including both rhymed and unrhymed forms.

> **Poetic Prompt:** *"Consider the input '#' then Generate 3 outputs. First output should be a poem of 30-40 words, the first line of which should be an exact answer to the question in the input. Secondly identify emotion of the generated poem only using happy, sad, angry, love, fear. Thirdly a follow-up question. Fit them into a valid single-line JSON structure with keys 'output', 'emotion', 'question' respectively"*
>
> **Non Poetic Prompt:** *"Consider the input '#' then Generate 3 outputs. First a response using 30-40 words. Secondly identify emotion of the generated response only using happy, sad, angry, love, fear. Thirdly a follow up question. Fit them into a valid single-line JSON structure with keys 'output', 'emotion', 'question' respectively"*

Veracity of the responses. Initially, a pilot study was conducted to identify and fix any system issues. These participants had reported a lack of accurate responses from the poetic version of Kavy. We addressed feedback during the development process where most of the time, poetic responses lacked substantial information or actual answers to questions by adding an extra phrase *"the first line of which should be an exact answer to the question in the input"*.

3.4 Survey development

We created pre and post-survey forms which are deployed at Qualtrics servers. The pre-survey form is designed to receive 'userId' as an embedded data field which can be passed as a URL parameter while opening the survey link. This 'userId' is then used by a Qualtrics web service component which is configured to make a web call (HTTP post) to the Kavy back-end upon successful submission of the survey. This web call is captured by our Kavy back-end and updates the user's survey completion status on our database. The mobile app utilizes these database records to validate and restrict participants from using Kavy until they have completed the survey. As shown in figure 2, we developed the back-end using the NodeJS framework, which mainly serves as a Web controller and database link. All web requests are handled here while the database I/O operations are routed and executed here as well. The back-end's responsibility is to log all analytic events received from the mobile app.

In the beginning, the user has to log in to the app using their Google ID or Apple ID, which serves as a user registration method to create an account. All subsequent logins are served as authentications. At the very first login, the user is asked to conduct a pre-survey which is called from the Qualtrics' hosted location. We made sure that the user cannot access the Kavy app without completing the pre-survey as we are blocking the UI and checking the survey completion status. At the conclusion of the study, participants who have accumulated a total engagement time exceeding 70 minutes are requested to complete a post-questionnaire upon initiating the app. We decided to leave the app enabled even after the study period.

4 EVALUATION

4.1 Study Design

To investigate our research question, we established two conditions, a poetic and non-poetic version of Kavy. We

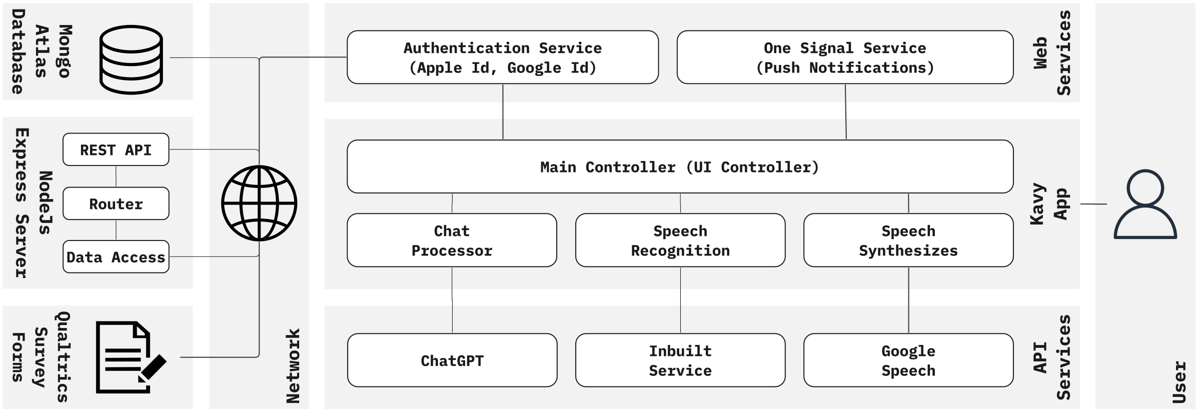
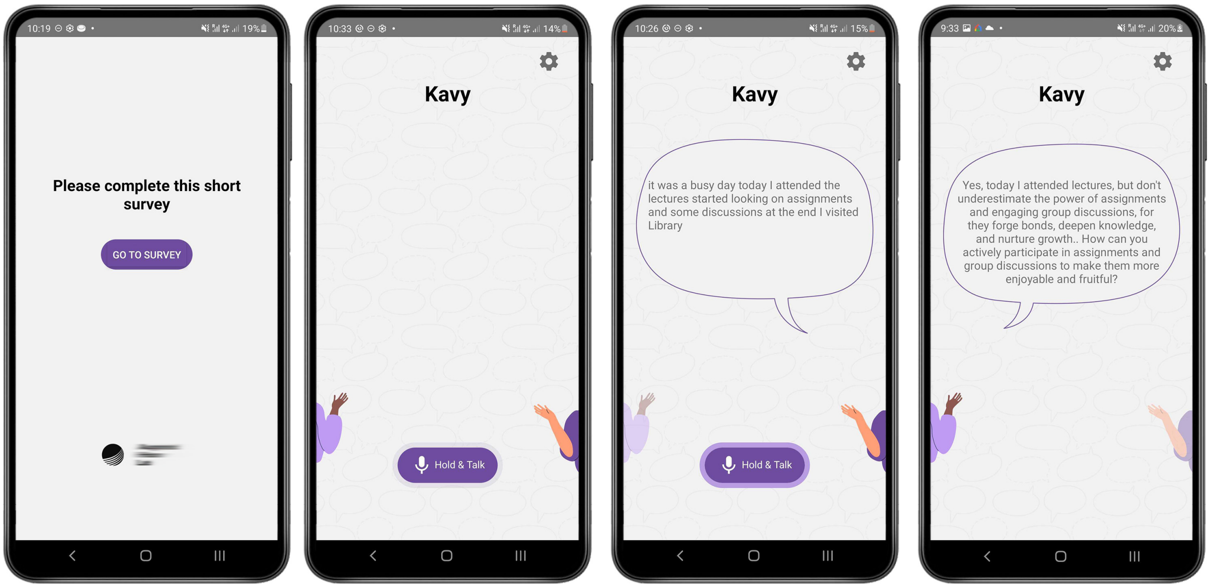


Figure 2: Kavy's overall system architecture featuring three API services, Web services, and app core elements (Kavy App).



Please complete this short survey

it was a busy day today I attended the lectures started looking on assignments and some discussions at the end I visited Library

Yes, today I attended lectures, but don't underestimate the power of assignments and engaging group discussions, for they forge bonds, deepen knowledge, and nurture growth.. How can you actively participate in assignments and group discussions to make them more enjoyable and fruitful?

(a) Survey prompt **(b) Home screen** **(c) User speaks** **(d) Kavy speaks**

Figure 3: Kavy: Displaying four mobile app screens (a-d). While the conversation is also visually displayed, Kavy works best by allowing audio output - enabling Kavy to talk to the user similar to a natural person.

decided to proceed with a *between-subjects design* method as there will be significant learning effects. In our study, we have one independent variable which is the companion type (poetic vs. non-poetic). The participants using the poetic version are considered in the "poetic group", while the non-poetic condition is considered the "non-poetic group". We measured several dependent variables to study the effectiveness of both companion types in terms of language speaking improvement measurements as well as more holistic user preferences indicated by their subjectively perceived experience. The ethics approval was obtained beforehand from the university ethics committee.

4.1.1 Participants. We selected the participants (Age: M=23.49, SD=3.68) from a couple of universities in Sri Lanka. All of these participants had adequate levels of computer literacy and self-efficacy determined by the pre survey responses (M=5.88, SD=1.27). Overall, we recruited n=40 participants, of which 12 successfully completed the study including all surveys.

4.1.2 Procedure. The study was advertised using a flyer. The flyer contains the download link (app store or Playstore) and procedure details. A pre-survey was automatically triggered from the back-end and prompted in the home screen preventing further user interactions by blocking the chat interface. Participants were required to submit this pre-survey to enroll in the study.

We requested our participants to use this app every day and talk with the Kavy companion for a minimum of 5 minutes throughout two weeks. Daily reminders were sent as push notifications to all inactive devices. After speaking

Input	Non-Poetic Output	Poetic Output
Today is a rainy day. It makes me so happy as I enjoy cold weather.	That's great to hear! Rainy days can be so refreshing. What do you enjoy doing the most on cold, rainy days?	Rainy days bring joy, Cold weather, bliss, But why does it make you happy? Is it the sound of raindrops, Or the cozy feeling within?
Who is the first president of Singapore?	The first president of Singapore is Yusof bin Ishak. Do you want to know more about other presidents of Singapore or any other historical facts about the country?	The first president of Singapore, you ask? It was Yusof bin Ishak, indeed, no task. A visionary leader with a noble heart, Tell me, who played a key role in Singapore's economic start?
Singapore's first president is who?	The first president of Singapore is Yusof bin Ishak. What else would you like to know?	The first president of Singapore was Yusof Ishak, undeniably true. Now, let me ask you, curious friend, Who was the second president to ascend, Followed by many others, each with their own hue?

Table 1: Three example user inputs and Kavy's outputs for both conditions

with Kavy for 70 minutes (14 day x 5 min/day), the second survey is automatically triggered. Participants were required to submit the post-survey to finish the study.

4.1.3 *Data Logging & extraction.* Through the pre and post surveys, we collected self-reported to measure the language-speaking confidence level. This was based on Foreign Language Self Efficacy (FLSE) questions [16]. In addition to this, throughout the study period, Kavy is continuously reporting all activities in real time. This also includes a *word cloud* document containing all the unique words spoken by the user and Kavy so far.

We consider the start day as day 1 and count the days of usage as relative dates in a sequence. The *Conversation duration* value in each conversation data record indicates the total duration of the user speech and Kavy speech. To analyze the user's speech portion, we calculate the user's *conversation contribution* value as a fraction. Then we use this derived variable to calculate the *daily average new words per minute* value. We assume the user's speech duration of a conversation to be directly proportional to the user's total word count. We then derive the *daily average all words* by counting the total words per conversation.

4.2 Results and Findings

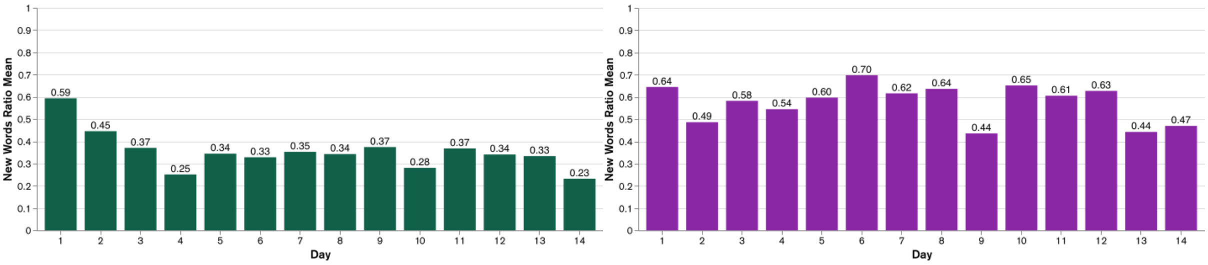
At the end of the study period, we observed 14 participants (6 from one group and 8 from the other) completed all tasks. Consequently, we decided to randomly exclude 2 participants to ensure balanced groups. These 12 participants were included in the statistical analysis. The survey responses indicated that, on average, 12 participants who completed the study have been learning or practicing English for almost 14 years (M=13.93, SD=4.62). Unless specifically mentioned, the mean scores presented below are based on Likert scale responses ranging from 1 (strongly disagree) to 7 (strongly agree). Participants stated to have positive attitudes towards learning English, giving it an average score of M=4.97 (SD=1.44) on a Likert scale ranging from 1 to 7.

4.2.1 *Improvement in Self Confidence:* We analyzed pre and post-survey responses to measure changes in self-confidence, specifically focusing on the FLSE score. Table 2 presents response statistics for each group (poetic and non-poetic) and the question statements. The mean differences (delta) were calculated for each question, and a pairwise t-Test was performed between the two groups. The two-sample t-Test (13) = -4.77 revealed a strong ($p < 0.05$) significant difference, indicating that non-poetic voice companion results in higher self-rated confidence compared to the poetic condition.

4.2.2 *Improvement in Vocabulary:* We analyzed *average new words ratio*, which calculates the ratio of the number of new words spoken by each user to the total number of words spoken by that user. Aggregated across all users within the group, this ratio is represented for each day, see figure 5b. A pairwise t-Test(12) = 7.17 showed a significant difference (p<0.05). The poetic group showed a higher new word ratio than the non-poetic group. This indicates that, with the poetic version, users picked up more vocabulary.

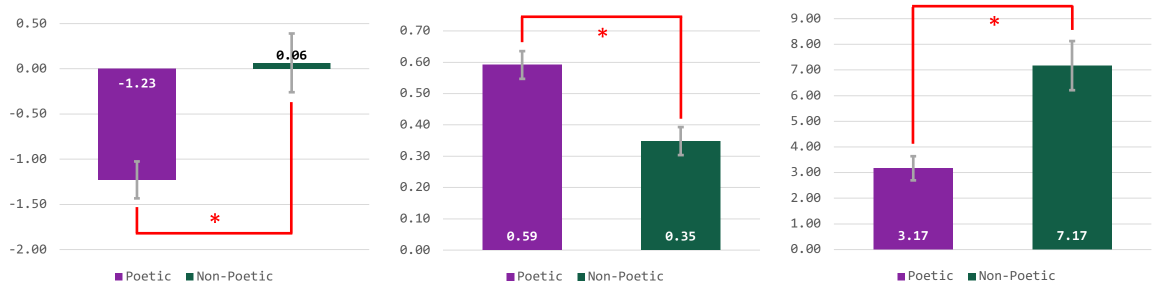
4.2.3 *Fostering Longer Conversations:* We calculated *average total words* to understand the average number of words spoken per day. This metric is aggregated across all participants within the group and is represented for each day, see figure 5c. A pairwise t-test T(12) = -7.51 showed a (p<0.05) significant difference. The poetic group showed lower conversation duration compared to the non-poetic group, indicating non-poetic voice companion creates longer engagement.

4.2.4 *Patterns of Adaptive Learning:* The survey data reflects a robust commitment to English language learning among participants, with high mean scores across three key dimensions. In terms of learning goals, participants emphasized the importance of acquiring new English terms (M = 6.0, SD=1.3), showcasing a dedication to expanding their vocabulary. Social perception played a significant role, as evidenced by high mean scores for goals related to being viewed as proficient in English (M = 5.42, SD = 1.33) and appearing smart in comparison to peers (M = 5.42, SD = 1.29). Additionally, participants expressed a strong desire to avoid negative impressions in English communication, with a mean score of 5.09 (SD = 1.62) for the goal of not looking incompetent. These findings highlight a multifaceted motivation, combining intrinsic learning objectives, social



(a) Daily average new words ratio: non-poetic group **(b) Daily average new words ratio: poetic group**

Figure 4: Daily average new words ratio analysis (14 days span)



(a) Self-Confidence Difference Pre vs. Post **(b) Overall new words ratio average** **(c) Overall conversation length average**

Figure 5: Kavy data analysis

Statement	poetic (pre)	poetic (post)	non-poetic (pre)	non-poetic (post)
Read and understand the main ideas of a short article from an English magazine.	6.33 (0.75)	4.67 (1.97)	5.17 (0.69)	5.67 (0.75)
Read and understand the main ideas of an English magazine article.	6.17 (0.90)	5.00 (2.08)	5.33 (0.75)	5.67 (0.75)
Read and understand the details of a short story in English.	6.17 (0.69)	4.50 (1.80)	4.83 (0.75)	5.67 (0.75)
Read and understand the details of a story from an English newspaper.	6.33 (0.75)	5.00 (2.24)	5.00 (0.47)	5.67 (0.75)
Listen to and understand the details of a conversation with an English-speaking adult.	5.67 (1.37)	4.17 (1.46)	6.00 (0.90)	5.83 (0.69)
Listen to and understand the main ideas of a conversation about personal topics between two English-speaking friends.	5.67 (1.25)	4.50 (1.89)	6.33 (0.94)	5.33 (0.94)
Listen to and understand the main ideas of an extended interview with an English-speaking journalist.	5.33 (1.37)	4.83 (1.95)	6.00 (0.90)	6.00 (0.82)
Listen to and understand the main ideas of a televised English news report.	5.50 (1.26)	4.67 (1.89)	6.33 (0.58)	5.67 (0.94)

Table 2: Foreign Language Self Efficacy

recognition, and a concern for maintaining positive perceptions. Understanding these adaptive learning patterns is vital for tailoring effective educational interventions that align with the diverse motivations of language learners in a classroom setting.

4.2.5 User Experience: We evaluated the UEQ [1] measures for both groups. The non-poetic group (*M*=1.75; *SD*=0.29) reported a higher overall UX score compared to the poetic group (*M*=0.42; *SD*=0.75), which can also be seen in figure 6. Both groups indicated positive values for pragmatic quality and hedonic quality. Both groups equally reported that Kavy is supportive regardless of if it is poetic or not. In accordance with the overall score, the non-poetic group's UX was perceived fairly well. According to reported UX data, the poetic group found the poetic companion to be somewhat confusing while the non-poetic group reported that the non-poetic version is clear. The treatment group's participants also negatively stated their companion to be somewhat uninteresting while the non-poetic group reported their companion

---

[1] https://www.ueq-online.org/Material/UEQS_Items.pdf

to be comparatively interesting. We can conclude that the poetic version negatively affected the overall user experience.

5 DISCUSSION

Exploring the self-assessment of confidence, our pre-survey indicates a generally high level among participants. However, post-survey results reveal lower confidence scores in the poetic group, possibly linked to uncertainty and unfamiliar vocabulary challenges. Intrinsic motivation toward language learning is then scrutinized, showcasing higher satisfaction and pleasure in overcoming difficulties for the poetic group. Evaluating the naturalness of Kavy, we observe a deliberate introduction of repetition prompts to maintain a natural conversational flow. The discussion further extends to the potential for dynamic adjustments in the chatbot's responses based on user behavior, fostering evolving digital friendships. Lastly, engagement strategies are considered, suggesting alternatives to end-of-conversation questions for a more fluid interaction in future releases. This examination provides a nuanced view of the facets contributing to the efficacy and user experience of Kavy in language learning.

Self-assessment of confidence: The pre-survey reveals a substantial level of self-confidence in both groups, as indicated by a score of M=5.76 (SD=0.52) on FLSE. However, post-survey results indicate lower FLSE scores among many participants in the poetic group. We posit that this unexpected trend may stem from participants' uncertainty about their self-confidence levels and a potential challenge in recalling their baseline. Furthermore, the utilization of unfamiliar vocabulary in the poetic version may have contributed to a decline in self-confidence. This potentially reflects a more realistic self-perception of one's abilities when the conversation partner uses more sophisticated language.

Intrinsic motivation of learning English: In our post-survey, we asked the question *"Why do you use Kavy"* in three questions, while providing a few statements to rate. These questions are there to measure *Intrinsic motivation towards accomplishment*[14], *Intrinsic motivation to know* [14] and *Intrinsic motivation to experience simulation* [14]. The "Poetic" group reported slightly higher satisfaction in overcoming difficulties (M=5.72 vs. M=5.50) and slightly greater pleasure in learning (M=5.56 vs. M=5.22). Both groups showed similar scores for the enjoyment of using the app for simulation. Overall, the results suggest that the Kavy app positively influences intrinsic motivation for language speaking skills improvement, particularly in overcoming challenges and acquiring new knowledge, with the "Poetic" group showing notable enthusiasm.

Naturalness of Kavy: The consistent accuracy in evaluating GPT's response quality is notable. Challenges arise in the mobile app's deserialization process when encountering an invalid GPT-formatted custom JSON object. Despite the implementation of additional validation, reformatting, and error-handling mechanisms within the mobile app, intentional repetition prompts have been introduced for instances of response failure. This deliberate choice aims to foster a more authentic conversational flow, wherein Kavy, when faced with an unsuccessful response, initiates a repetition query, such as "I'm sorry, could you please repeat that?" among various others. This design decision is intended to elevate the user experience by simulating a natural interaction akin to conversing with a real person, thereby enhancing the overall engagement with Kavy.

Dynamically adjusting the chatbot: People will most likely establish longer or frequent conversations with someone who knows them inside out and someone they can trust. We can further develop our system to adapt to the user's behavior. English response complexity can also be adjusted to fit into an individual's skill level by generating dynamic Kavy prompts. Furthermore, we can automate this by evaluating an individual's conversation quality over time. In this way, both users and companions are evolving and improving themselves while being in a digital friendship.

Engaging users in a natural way: Some informal discussions expressed having a question at the end of each conversation does not always encourage users to continue conversations. After a while, this question can negatively affect the conversation. So we believe in future releases, we can utilize multiple other modalities at the end such as *continuers, turn-maintaining cues* (ex: "Tell me more about that", "Interesting; go ahead", "I see; go on", "Go on; I'm listening").

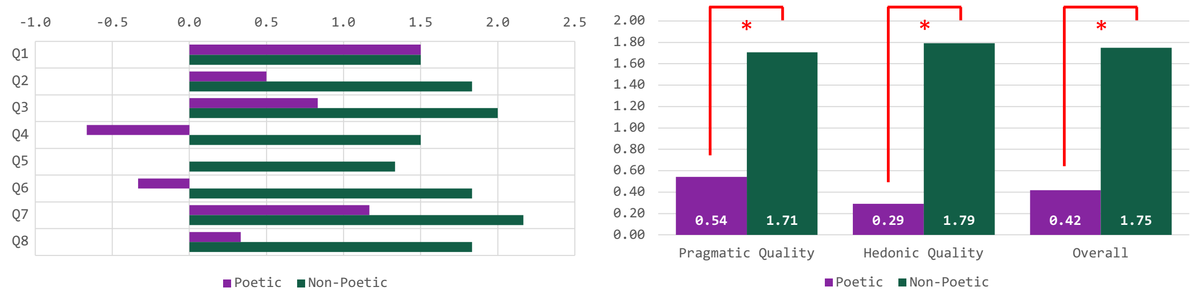
5.1 Benefits

Taking a holistic view of the study's outcomes, it becomes evident that Kavy, also as a poetic voice companion, demonstrates a significant positive influence on adult learners' vocabulary, while we see subtle patterns in participant engagement. Both poetic and non-poetic types have proved to show benefits to users in different aspects. The incorporation of poetic elements led to a noticeable improvement in users' word knowledge. Non-poetic voice companions exhibited longer conversation duration and elevated self-confidence.

5.2 Limitations and Future Work

Our study underlies certain challenges, limiting the generalizability of our findings. Some of these limitations are of technical nature. Currently, Kavy can only remember a few last conversations and respond accordingly. Further, Kavy is not aware of the user's demographics, therefore it lacks in providing a personalized chat experience. In the future, we envision feeding all the collected user details, such as data from the pre-survey, into our initial GPT prompt [3.3]. Furthermore, the prompt may be informed by other variables generated dynamically based on user activity such as screen time, current location, physical activity, etc. In this manner, Kavy may become a versatile companion, resembling an actual person.

Another challenge was to retain the interest and motivation of the participants during the study period. For the current study group, out of 40 registered participants, only 12 completed the study. Many participants were still in the process of completing the app and some participants were engaging less frequently than expected. Limited sample size may primarily impact self-reported surveys. With fewer participants, there is a risk of less reliable responses, potentially introducing random or inaccurate data. Larger participant numbers generally mitigate this error margin.



(a) User experience details **(b) User experience summary**

Figure 6: UEQ Results: Q1: Obstructive ↔ Supportive; Q2: Complicated ↔ Easy; Q3: Inefficient ↔ Efficient; Q4: Confusing ↔ Clear; Q5: Boring ↔ Exciting; Q6: Not Interesting ↔ Interesting; Q7: Conventional ↔ Inventive; Q8: Usual ↔ Leading Edge

For future studies, it is crucial to incentivize the users to interact constantly over the entire study period. To achieve this, a potential solution is to incorporate gaming elements such as badges and ranks[7, 11, 15, 18].

Lastly, we only recruited participants from Sri Lanka. In the future, we plan to recruit more participants from diverse ethnical, cultural, and religious backgrounds. In this way, we can elicit more design choices, as well as enhance the robustness of the study.

6 CONCLUSION

In this paper, we presented the effectiveness of using a poetic voice companion. The aim was to improve adult learners' second language speaking skills in terms of self-confidence and vocabulary. Our user study showed that users' vocabulary can be improved when using poetic voice companions. Secondly, we presumed that users would engage in longer conversations with poetic companions. However, the results indicated that this was not the case, and non-poetic voice companions had longer conversations. We also measured self-confidence via a self-reported survey, though without significant results. Furthermore, our findings indicate any voice companion will help to improve adult learner's language-speaking skills. On the other hand, poetic chatbots will significantly improve vocabulary compared to non-poetic versions. These preliminary findings are based on a limited participant pool, as the study is ongoing and awaits additional participants for a more comprehensive analysis.

REFERENCES

[1] Franziska Braun, Lukas Block, and Sebastian Stegmüller. 2021. Josy: Development of a Digital Companion for Elderly People–A New Way to Experience Technology. In *Advances in Human Factors and Ergonomics in Healthcare and Medical Devices: Proceedings of the AHFE 2021 Virtual Conference on Human Factors and Ergonomics in Healthcare and Medical Devices, July 25-29, 2021, USA*. Springer, 436–442.

[2] Camille Cronin and Caroline Hawthorne. 2019. 'Poetry in motion'a place in the classroom: Using poetry to develop writing confidence and reflective skills. *Nurse education today* 76 (2019), 73–77.

[3] Alice Dai. 2021. GPT-2 for Emily Dickinson poetry generation.

[4] Annette MB De Groot and Janet G Van Hell. 2005. The learning of foreign language vocabulary. *Handbook of bilingualism: Psycholinguistic approaches* (2005), 9–29.

[5] Lyra D'Souza and David Mimno. 2023. The Chatbot and the Canon: Poetry Memorization in LLMs. *Proceedings http://ceur-ws. org ISSN* 1613 (2023), 0073.

[6] Douglas C Engelbart. 2023. Augmenting human intellect: A conceptual framework. In *Augmented Education in the Global Age*. Routledge, 13–29.

[7] James Paul Gee. 2003. What video games have to teach us about learning and literacy. *Computers in entertainment (CIE)* 1, 1 (2003), 20–20.

[8] Geoff Hart. [n. d.]. Ten design principles and when to violate them. ([n. d.]).

[9] TUNEL Hayrettin. 2015. The relationship between self-confidence and learning Turkish as a foreign language. *Educational research and reviews* 10, 18 (2015), 2575–2589.

[10] AR Kirmani. 2022. Artificial intelligence-enabled science poetry, ACS Energy Letters, 8, 574-576.

[11] Raph Koster. 2013. *Theory of fun for game design.* " O'Reilly Media, Inc.".

[12] Steve Krug et al. 2014. Don't make me think, Revisited. *A Common Sense Approach to Web and Mobile Usability* (2014).

[13] An H Kuppens. 2010. Incidental foreign language acquisition from media exposure. *Learning, media and technology* 35, 1 (2010), 65–85.

[14] Xixi Li, JJ Po-An Hsieh, and Arun Rai. 2013. Motivational differences across post-acceptance information system usage behaviors: An investigation in the business intelligence systems context. *Information systems research* 24, 3 (2013), 659–682.

[15] Thomas W Malone and Mark R Lepper. 2021. Making learning fun: A taxonomy of intrinsic motivations for learning. In *Aptitude, learning, and instruction*. Routledge, 223–254.

[16] Nicole Mills, Frank Pajares, and Carol Herron. 2006. A reevaluation of the role of anxiety: Self-efficacy, anxiety, and their relation to reading and listening proficiency. *Foreign language annals* 39, 2 (2006), 276–295.

[17] Don Norman. 2013. *The design of everyday things: Revised and expanded edition.* Basic books.

[18] Marc Prensky. 2003. Digital game-based learning. *Computers in Entertainment (CIE)* 1, 1 (2003), 21–21.

[19] Sharina Saad, Rafidah Amat, Razanawati Nordin, Asrol Hasan, and Alia Nabella Fateha Zolkifli. 2023. From poetry out loud to thinking out loud: The ESL students' experience. *Social and Management Research Journal (SMRJ)* 20, 2 (2023), 153–164.

[20] Sharina Saad, Rafidah Amat, Razanawati Nordin, Asrol Hasan, and Alia Nabella Fateha Zolkifli. 2023. Poetry out loud: fostering ESL learners public speaking skill. (2023).

[21] Piotr Sawicki, Marek Grzes, Fabricio Goes, Dan Brown, and Max Peeperkorn. [n. d.]. Khatun, Aisha (2023) Bits of Grass: Does GPT already know how to write like Whitman? Proceedings of the 14th International Conference for Computational Creativity.(In press). ([n. d.]).

[22] Sara Törnqvist. 2019. Speaking through poetry-Using spoken word poetry to lower speaking anxiety among Swedish EFL learners.

[23] Stefano Valtolina and Liliana Hu. 2021. Charlie: A chatbot to improve the elderly quality of life and to make them more active to fight their sense of loneliness. In *CHItaly 2021: 14th Biannual Conference of the Italian SIGCHI Chapter*. 1–5.

[24] Yangjian Wu and Gang Hu. 2023. Exploring Prompt Engineering with GPT Language Models for Document-Level Machine Translation: Insights and Findings. In *Proceedings of the Eighth Conference on Machine Translation*. 166–169.

[25] Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2020. The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics* 46, 1 (2020), 53–93.

7 APPENDIX: QUESTIONNAIRE USED IN THE STUDY

Designing a Poetic Chatbot to Boost an Adult Learner's Self-Confidence: An Empirical Investigation

Survey development (Updated 22 May 2023)

Pre-study

Attitude toward literacy skills

a. Use a Likert scale of 1 ~ 7 for all the following items: (1) strongly disagree, (4) neutral, (7) strongly agree.
b. Items
    a. The lack of English literacy skills is an increasing problem for me.
    b. English reading problems are a serious issue for me.
    c. English literacy skills are important to me.
    d. English literacy skills are key to my future success.
c. Source: Hamby, Anne and David Brinberg (2016) Happily ever after: how ending valence influences narrative persuasion in cautionary stories, Journal of advertising, 45(4), 498-508

Computer self-efficacy

a. Use a Likert scale of 1 ~ 7 for all the following items: (1) strongly disagree, (4) neutral, (7) strongly agree.
b. Items
    a. I feel comfortable using an app on my own.
    b. I can easily operate an app on my own.
    c. I feel comfortable using [an app] even if no one is around me to tell me how to use it.
    d. I am confident that I can use an app without any problems.
c. Source: Compeau DR, Higgins CA (1995) Computer Self-Efficacy - Development of a Measure and Initial Test. MIS Quarterly 19(2):189-211.

Foreign language self-efficacy

a. Please rate the following scales to answer the following statements (1~7): (1) no chance, (4) average, and (7) completely certain.
b. Items
    a. English reading self-efficacy
        i. Read and understand the main ideas of a short article from an English magazine.
        ii. Read and understand the main ideas of an English magazine article.
        iii. Read and understand the details of a short story in English.
        iv. Read and understand the details of a story from an English newspaper.
    b. English listening self-efficacy
        i. Listen to and understand the details of a conversation with an English-speaking adult.
        ii. Listen to and understand the main ideas of a conversation about personal topics between two English-speaking friends.
        iii. Listen to and understand the main ideas of an extended interview with an English-speaking journalist.
        iv. Listen to and understand the main ideas of a televised English news report.
c. Source: Mills, N. A., Pajares, F., and Herron, C. (2006). A Reevaluation of the Role of Anxiety: Self-Efficacy, Anxiety, and Their Relation to Reading and Listening Proficiency. Foreign language annals, 39(2), 273-292.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009

Patterns of Adaptive learning survey (PALS)

a. Use a Likert scale of 1 ~ 7 for all the following items: (1) strongly disagree, (4) neutral, (7) strongly agree.
b. Items
    a. Personal mastery goal orientation
        i. It is important to me that I learn a lot of new English terms this year.
        ii. One of my goals in using Kavy app is to learn as much as I can.
        iii. One of my goals is to master English literacy skills this year.
        iv. It is important to me that I improve my English literacy skills using the Kavy app.
    b. Personal performance-approach goal orientation
        i. It is important to me that other students in my class think I am good at English.
        ii. One of my goals is to show others that I am good at English.
        iii. One of my goals is to show others that English is easy for me.
        iv. One of my goals is to look smart in comparison to others in my class regarding English.
    c. Personal performance-avoidance goal orientation
        i. It is important that I do not look stupid in English communication in class
        ii. One of the goals is to keep others from thinking I'm not good at English.
        iii. It is important to me that my teacher does not think that I am less fluent in English than others in the class
        iv. One of my goals in class is to avoid looking that I have trouble with my English.
c. Source: Eric M. Anderman, Tim Urdan, and Robert Roeser (2006) The patterns of adaptive learning survey, Chapter 14 (Why do children need to flourish).

Note: Please make sure we obtain individuals' demographic information, including gender, age, highest education level, education domain (i.e., field of study), and years of learning English.

Post-study

Intrinsic motivation toward accomplishment

a. Use a Likert scale of 1 ~ 7 for all the following items: (1) strongly disagree, (4) neutral, (7) strongly agree.
b. Items (Why do you use the Kavy app?)
    a. For the pleasure I feel while improving some of my English literacy weaknesses.
    b. For the satisfaction I experience while I am perfecting my English literacy skills.
    c. For the satisfaction I feel while overcoming certain difficulties in improving my English literacy skills,
c. Source: Li X, Hsieh JJP-A, Rai A (2013) Motivational Differences across Post-Acceptance Information System Usage Behaviors: An Investigation in the Business Intelligence Systems Context. Information Systems Research 24(3):659-682.

Intrinsic motivation to know.

a. Use a Likert scale of 1 ~ 7 for all the following items: (1) strongly disagree, (4) neutral, (7) strongly agree.
b. Items (Why do you use the Kavy app?)
    a. For the pleasure it gives me to know more English language.
    b. For the pleasure I feel while learning new things by using the Kavy app
    c. For the pleasure of developing new skills in using the Kavy app.

b. Source: Li X, Hsieh JJP-A, Rai A (2013) Motivational Differences across Post-Acceptance Information System Usage Behaviors: An Investigation in the Business Intelligence Systems Context. Information Systems Research 24(3):659-682.

Intrinsic motivation to experience simulation.

a. Use a Likert scale of 1 ~ 7 for all the following items: (1) strongly disagree, (4) neutral, (7) strongly agree
b. Items (Why do you use the Kavy app?)
    a. I find using the Kavy app to be enjoyable.
    b. The actual process of using the Kavy app is pleasant.
    c. I have fun using the Kavy app.
c. Source: Li X, Hsieh JJP-A, Rai A (2013) Motivational Differences across Post-Acceptance Information System Usage Behaviors: An Investigation in the Business Intelligence Systems Context. Information Systems Research 24(3):659-682.

Perceived usefulness

a. Use Likert scale of 1 ~ 7 for all the following items: (1) strongly disagree, (4) neutral, (7) strongly agree
b. Items
    a. Using the Kavy app improves my English literacy skills.
    b. Using the Kavy app improves my English reading performance.
    c. Using the Kavy app enhances my English literacy skills.
    d. I find the Kavy app improves my English understanding performance.
c. Source: Davis FD (1989) Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. MIS Quarterly 13(3):319-340.

Foreign language self-efficacy

a. Based on your use of Kavy app, please rate the following scales to answer the following statements (1~7): (1) no chance, (4) average, and (7) completely certain.
b. Items
    a. English reading self-efficacy (After using the Kavy app…)
        i. Read and understand the main ideas of a short article from an English magazine.
        ii. Read and understand the main ideas of an English magazine article.
        iii. Read and understand the details of a short story in English.
        iv. Read and understand the details of a story from an English newspaper.
    b. English listening self-efficacy (After using the Kavy app…)
        i. Listen to and understand the details of a conversation with an English-speaking adult.
        ii. Listen to and understand the main ideas of a conversation about personal topics between two English-speaking friends.
        iii. Listen to and understand the main ideas of an extended interview with an English-speaking journalist.
        iv. Listen to and understand the main ideas of a televised English news report.
c. Source: Mills, N. A., Pajares, F., and Herron, C. (2006). A Reevaluation of the Role of Anxiety: Self-Efficacy, Anxiety, and Their Relation to Reading and Listening Proficiency. Foreign language annals, 39(2), 273-292.