# DESIGNING A POETIC VOICE COMPANION TO IMPROVE ADULT LEARNER'S SELF-CONFIDENCE: KAVY

B S Cooray

238012H

Degree of Master of Science (With a major component of research)

Department of Computer Science and Engineering
Faculty of Engineering

University of Moratuwa
Sri Lanka

Feb 2024

# DESIGNING A POETIC VOICE COMPANION TO IMPROVE ADULT LEARNER'S SELF-CONFIDENCE: KAVY

B S Cooray

238012H

Thesis submitted in partial fulfillment of the requirements for the degree
Degree of Master of Science (With a major component of research)

Department of Computer Science and Engineering
Faculty of Engineering

University of Moratuwa
Sri Lanka

Feb 2024

## DECLARATION

I declare that this is my own work and this Thesis does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or Institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature: Date:

The supervisors should certify the Thesis with the following declaration.

The above candidate has carried out research for the Degree of Master of Science (With a major component of research) Thesis under our supervision. We confirm that the declaration made above by the student is true and correct.

Name of Supervisor: Dr. Chathuranga Hettiarachchi

Signature of the Supervisor: Date:

Name of Supervisor: Ms. Vishaka Nanayakkara

Signature of the Supervisor: Date:

Name of Supervisor: Prof. Suranga Nanayakkara

Signature of the Supervisor: Date:

## DEDICATION

This thesis is dedicated to my family and friends. To my parents, for their endless love, support, and sacrifices that have allowed me to reach this milestone. To my mentors and professors, whose guidance and wisdom have inspired and challenged me. To my friends, who have motivated and helped me in various aspects of everyday life, especially in my academic journey. And to all those who believed in me and supported me along the way, thank you for being by my side. This work represents one small step toward the successful journey ahead.

## ACKNOWLEDGEMENT

This research was funded by the Augmented Human Lab at the National University of Singapore. I extend my deepest gratitude to my supervisors, A/Prof Suranga Nanayakkara from the National University of Singapore, Dr. Chathuranga Hettiarachchi from the University of Moratuwa, and Mrs. Vishaka Nanayakkara from the University of Moratuwa, for their invaluable support and guidance throughout the research and user studies. Suranga closely supervised all aspects of my work as I conducted my research at his lab, the Augmented Human Lab. Dr. Chathuranga provided extensive feedback as my most immediate supervisor, managing weekly online meetings and all university arrangements. Mrs. Vishaka encouraged me from the outset and played a pivotal role in initiating my master's enrollment.

I also acknowledge the tremendous support provided by my colleague, Yasith Samaradivakara, who developed the early version of the proof-of-concept app.

I am grateful to all my colleagues at the Augmented Human Lab for their participation in the pilot study, their constructive feedback, and the fruitful discussions that led to the engineering of the prompts. Your contributions have been instrumental in the completion of this work.

Additionally, I would like to thank Denys Matthies for his contributions to the research paper titled "Kavy: Fostering Language Speaking Skills and Self-Confidence Through Conversational AI," which was published at the Augmented Humans 2024 conference.

Finally, I extend my thanks to everyone who has supported this research in any capacity, even if not mentioned by name. Your help and encouragement are deeply appreciated.

## ABSTRACT

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.

A conversational chatbot, Kavy, was developed as a companion to help these individuals practice English. In a study with 40 users, the effectiveness of Kavy in improving communicators' proficiency (e.g., verbal expression, conversation length, quality of speech) and self-confidence was investigated, using both poetic and non-poetic conversational styles. It was found that users were highly motivated by the poetic version, resulting in a significant increase in vocabulary. However, the poetic chatbot also faced challenges, as it could be confusing when the learner's level of English was rather low.

This pioneering work is seen as a first and promising approach that should continue to be investigated in the future.



**Keywords**: Social Chatbots, Conversational AI Agents, Voice Interfaces, Artificial Intelligence, Language Studies, Self Confidence, Poetry, Cognitive Augmentation

# TABLE OF CONTENTS

## LIST OF FIGURES

## LIST OF TABLES

## LIST OF ABBREVIATIONS

| **Abbreviation** | **Description** |
|---|---|
| AI | Artificial Intelligence |
| API | Application Programming Interface |
| FLSE | Foreign Language Self-Efficacy |
| GPT | Generative Pre-trained Transformer |
| JSON | JavaScript Object Notation |
| LLM | Large Language Models |
| MQL | Mongo Query Language |
| NodeJS | Node Java Script |
| NUS | National University of Singapore |
| QR | Quick Response |
| UI | User Interface |
| URL | Uniform Resource Locator |

## LIST OF APPENDICES

# CHAPTER 1

# INTRODUCTION

## 1.1 Overview

Learning a new language can be challenging and time-consuming. In general, most people may have good literacy in three components of language skills: reading, writing, and listening. However, beginners often struggle with the speaking component, lacking the ability to sustain a conversation due to insufficient vocabulary and self-confidence. This research introduces a novel approach to addressing these challenges through the development and evaluation of a mobile and personal voice companion named Kavy. This voice companion is designed to converse poetically with a human-like voice and can handle imperfect and immature user input.

## 1.2 Background

### 1.2.1 Current Language Learning Challenges

Language learners, especially beginners, often face significant challenges in developing their speaking skills. The lack of self-confidence, vocabulary, and exposure to conversational practice are major barriers to progress [1–3].

### 1.2.2 Role of Technology in Language Learning

Virtual voice companions, widely available and suitable for assisting humans in performing various tasks, hold the potential to support language learning. Practicing speaking in a virtual environment with a virtual assistant could 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. Additionally, current implementations of voice companions struggle with handling imperfect or incomplete speech, which hinders the ability of users to improve their language-speaking skills.

## 1.3 Motivation

### 1.3.1 Need for Improved Language Support

The primary motivation behind this research is to explore innovative methods for supporting adult English learners, particularly within university contexts. In Sri Lanka, numerous faculties deliver coursework in English, providing significant advantages to students whose mother tongue is not Sinhalese. However, many learners perform well in reading and writing while their listening and speaking skills lag behind. Despite the availability of additional English programs aimed at teaching academic English, these efforts often fall short of addressing this specific issue.

### 1.3.2 The Potential of AI and Mobile Technology

Given the ubiquity of mobile phones and the rising popularity of Artificial Intelligence (AI) there is a wealth of opportunities for leveraging modern technology to assist learners in overcoming language barriers. Social voice companions have the capacity to engage individuals in meaningful conversations. Thus, it is logical to explore the potential of utilizing a social companion as a tool to support language learning. This exploration leads to the design and development of a social voice companion tailored to adult learners, aiming to bolster their self-confidence in speaking English.

### 1.3.3 Limitations of Current Voice Companions

Modern voice companions often fall short of optimization for language learning. Primarily designed as assistant-type interfaces, they typically execute predefined actions or offer support in various tasks. However, many of these chatbots lack the flexibility to respond effectively to incomplete inputs from users. Moreover, some provide only brief or context-specific replies, potentially leading to premature termination of conversations and decreased user engagement.

### 1.3.4 Integration of Large Language Models (LLM)

To address these limitations, a novel approach is required—one that integrates Large Language Models (LLM) to create a specialized voice companion. Unlike conventional chatbots, this companion would be finely tuned to engage users more dynamically. Rather than offering standard responses, it would employ poetic replies, captivating users and fostering deeper engagement. By encouraging increased interaction and dialogue, this innovative approach aims to enhance users' self-confidence, while also improving their vocabulary and language-speaking skills.

## 1.4 Problem Statement

The problem statement arises from the observation that despite the prevalence of English-medium coursework in Sri Lankan universities, many adult learners struggle with listening and speaking skills, even if they excel in reading and writing. Existing English programs fail to adequately address this specific issue. Furthermore, modern voice companions, primarily designed as assistant-type interfaces, lack the flexibility to respond effectively to incomplete inputs from users and often provide brief or context-specific replies, leading to decreased user engagement. Therefore, there is a need for a specialized voice companion that integrates LLMs and employs dynamic engagement strategies, such as poetic replies, to enhance users' self-confidence and language-speaking skills.

## 1.5 Objectives

The objectives of this research encompass several key aspects aimed at addressing the identified challenges in language learning support for adult English learners:

- Understanding Linguistic Barriers: Develop a comprehensive understanding of the specific linguistic barriers faced by adult learners, particularly in listening and speaking skills within university contexts in Sri Lanka.
- Designing a Specialized Voice Companion: Design and develop a specialized voice companion tailored to the unique needs of adult learners.
- Integrating LLMs: Integrate LLMs into the companion's framework to dynamically adapt and respond to user inputs, fostering deeper engagement and enhancing learning outcomes.
- Employing Innovative Engagement Strategies: Explore innovative engagement strategies, such as employing poetic replies, to encourage increased interaction and dialogue between users and the voice companion.
- Evaluating Effectiveness: Rigorously evaluate the effectiveness of the developed companion in improving users' self-confidence and new vocabulary usage in language-speaking skills through user testing and feedback analysis.

## 1.6 Hypotheses

These are the proposed hypotheses to evaluate this research.

$H_1$ : A Poetic voice companion will increase self-confidence

$H_2$ : User's vocabulary will be improved regardless of companion type

$H_3$ : Poetic companions will engage the user in a longer conversation

## 1.7 Research Methodology Overview

### 1.7.1 Study Design

The study employs a between-subjects design, focusing on comparing two versions of the voice companion: poetic and non-poetic. This design helps isolate the effects of the type of companion on the dependent variables, ensuring that any differences observed are due to the companion type and not other extraneous factors.

### 1.7.2 Participants

Participants for this study were recruited from university settings, targeting adult learners who are interested in improving their English-speaking skills. Detailed inclusion and exclusion criteria ensured the selection of suitable participants.

## 1.8 Research Contribution

The primary contributions of this research to the HCI community are the Kavy software solution and an associated survey system. The software solution comprises three main components: a front-end mobile app, a back-end server, and a database service. In addition to these primary contributions, this research also presents several empirical findings derived from a user study. All of this research has been published in a paper titled *Kavy: Fostering Language Speaking Skills and Self-Confidence Through Conversational AI* at *Augmented Humans 2024* and is available in the *ACM Digital Library* [4].

## 1.9 Chapter Summary

- Chapter 1: Introduction - Provides an overview of the study, background, motivation, problem statement, objectives, hypotheses, and an outline of the thesis.
- Chapter 2: Research Method - Describes the research design, materials, procedures, data collection, and analysis plans.
- Chapter 3: Implementation - Details the development and technical aspects of the Kavy app.
- Chapter 4: Results and Analysis - Presents the findings of the study, including statistical analysis and user feedback.
- Chapter 5: Discussion - Discusses the implications of the findings, limitations, and future research directions.
- Chapter 6: Conclusion - Summarizes the study, its contributions, and final thoughts.

# CHAPTER 2

# LITERATURE REVIEW

## 2.1 Introduction

In this chapter, the existing literature on social chatbots, large language models (LLMs) for poem generation, the relationship between poetry and self-confidence, and voice-based language learning will be reviewed. Additionally, recent advancements in AI and conversational technologies will be discussed. This review provides a foundation for understanding the context and significance of developing Kavy, a poetic voice companion designed to enhance language learning and self-confidence in adult learners.

## 2.2 Social chatbots

There is a body of work exploring the use of Artificial Intelligence (AI) conversational chatbots to engage users in meaningful conversations [5, 6]. For instance, Kaluarachchi et al. investigated how a poetic voice companion could engage people in conversations with relevant output [7]. Microsoft developed an AI chatbot XiaoIce [8] (Figure 1, 2) that can assist with people's communication, affection, and social belonging. SimSimi [9] is a popular open-domain social chatbot that facilitates user engagement by allowing them to express emotions such as sadness and depression. These studies demonstrate the potential of social chatbots to foster user engagement and emotional connection in conversations.

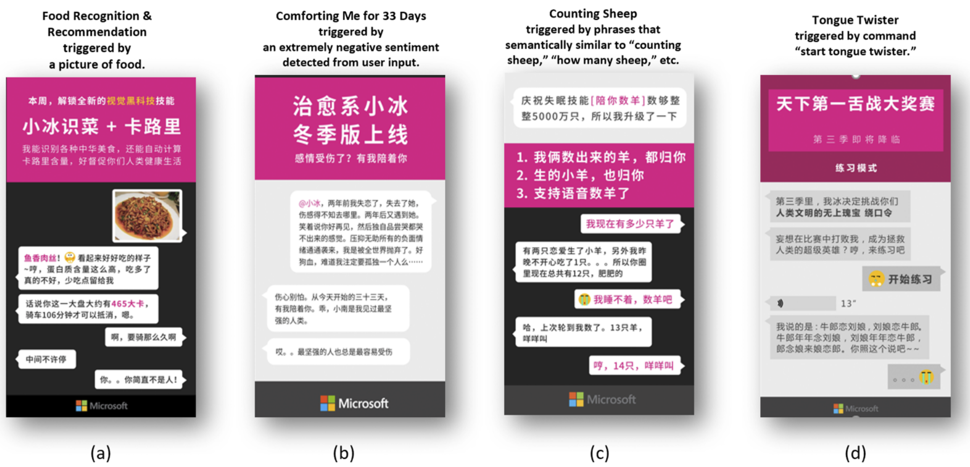



Fig. 2.1: Existing Apps: Xiaolce by Microsoft

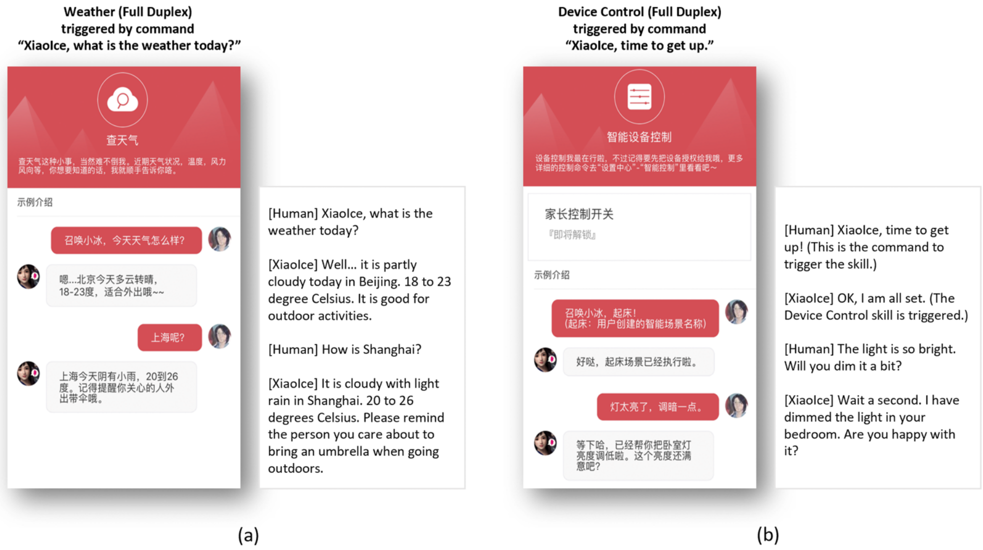



Fig. 2.2: Existing Apps: Xiaolce by Microsoft

## 2.3 Language Learning Chatbots

Studies have explored chatbots specifically designed for language learning. For example, Duolingo (Figure 3) introduced a chatbot feature that allows users to practice conversation in a simulated environment [10]. Valtolina et al. developed Charlie (Figure 4), a chatbot aimed at helping students learn English through conversation practice [5]. These chatbots provide immediate feedback and are accessible at any time, making them valuable tools for language learners. However, they often lack the depth of engagement and personalization needed to significantly boost confidence and conversational skills.

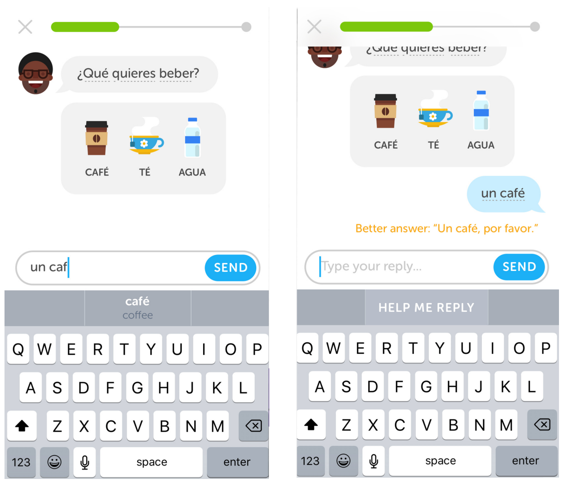



Fig. 2.3: Existing Apps: Duolingo chat interface

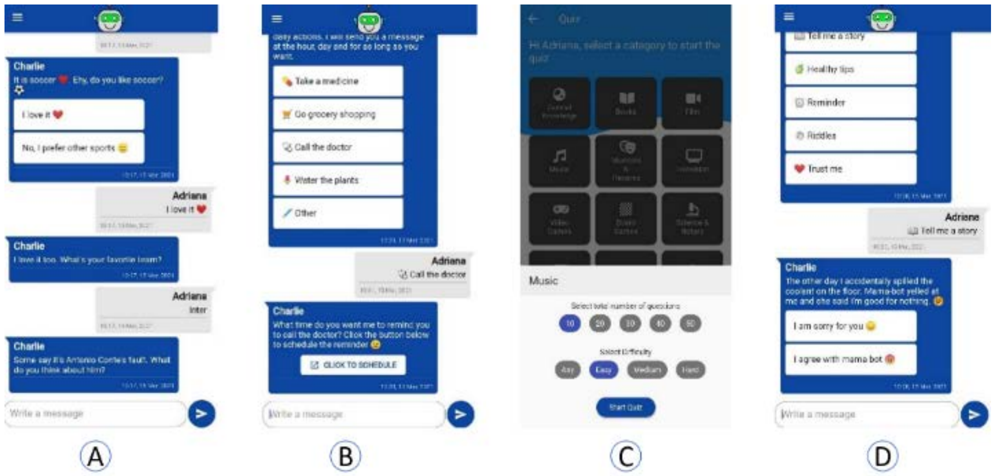



Fig. 2.4: Existing Apps: Charlie app interface

## 2.4 LLM poem generation

Numerous studies have proven the ability and power of large language models (LLMs) like ChatGPT to generate poems using specific prompts. Research indicates [11] that ChatGPT can memorize and retrieve famous poems as it is trained on web data. A. Dai et al. proposed a method that fine-tunes the Generative Pre-trained Transformer (GPT) 2 model to generate bespoke poems curated to match famous poet Emily Dickinson's writing style [12]. A similar study by Sawicki P. et al. demonstrated the usage of GPT models to generate poems in Walt Whitman's style [13]. A.R. Kirmani [14] documented a possible way to utilize GPT models to generate poems, especially auto-crafted ones in the science domain. Additionally, Ghazvininejad et al. [15] introduced a neural model for generating high-quality poems with controllable styles and themes. Furthermore, Lau et al. [16] developed a hierarchical neural model that generates poems by first planning the high-level structure and then filling in the details. Yan et al. [17] proposed a method for generating Chinese classical poetry using a two-stage approach with a planning model and a generation model. Zhang et al. [18] introduced a memory-augmented neural model that can generate poems by extracting and recombining lines from existing poems.

## 2.5 Poetry and self-confidence

Numerous studies have discovered that the nature of poetry correlates with human self-confidence in several ways [19]. Sara Törnqvist's study [20] found that individuals may feel less anxious speaking English when performing spoken word poetry. Furthermore, her research indicates that poems significantly impact students' motivation to perform. Poetry out loud [21] and Poetry out loud to thinking out loud [22] 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, reading fluency, and enhancing public speaking abilities. Additionally, Chandler et al. [23] demonstrated that engaging in creative writing activities like poetry can boost self-esteem and emotional well-being. Hijazi et al. [24] conducted a study showing that poetry therapy can significantly reduce anxiety and improve self-esteem among university students. Finch A. [25] explored the use of poetry in language learning and found that it can enhance students' motivation, self-expression, and cultural awareness. Furthermore, Simecek et al [26] investigated the role of poetry in developing language skills and personal growth, highlighting its potential for fostering creativity and self-confidence.

## 2.6 Voice-Based Language Learning

Several studies have explored the use of voice-based interfaces and conversational agents for language learning [27–29]. Huang, H. H et al. [29] developed a voice-based conversational agent for English language learning, which showed positive results in improving students' speaking skills. Katsarou et al. [30] proposed a voice-based language learning system using automatic speech recognition and natural language processing to provide personalized feedback, highlighting the potential of Intelligent Virtual Agents (IVAs) in enhancing English as a Foreign Language (EFL) learning. Petrovic et al. [31] investigated the use of voice-based conversational agents for practicing speaking skills in a foreign language, finding that learners generally had a positive experience and perceived the system as helpful. Additionally, Dousay [32] explored the use of voice-based virtual assistants like Amazon Alexa for language learning, highlighting their potential for providing authentic conversational practice and personalized feedback.

## 2.7 Conversational AI in Education

The application of conversational AI in education is expanding. Conversational agents are increasingly used in tutoring systems, providing personalized assistance to students. For instance, Graesser et al. developed AutoTutor, an intelligent tutoring system that engages students in natural language dialogues to facilitate learning [33]. AutoTutor has been shown to improve learning outcomes by providing immediate feedback and adapting to the learner's responses.

## 2.8 Emotional Engagement and AI

Emotional engagement is crucial for effective learning, and AI can play a significant role in enhancing this aspect. McDuff D. et all developed an AI that can detect and respond to human emotions, providing a more engaging and empathetic user experience [34]. Such technologies can be integrated into educational chatbots to make interactions more personalized and emotionally supportive.

## 2.9 Language Learning and Poetic Expression

Integrating poetic expression into language learning can make the process more enjoyable and engaging. Poetry can be used to teach various aspects of language, including rhythm, phonetics, and vocabulary. Studies have shown that using poetry in language instruction can enhance creativity and motivation among learners [35]. Poetic devices such as metaphor and simile can also help learners grasp complex language concepts more easily.

## 2.10 Personalized Learning Experiences

Personalized learning experiences are becoming increasingly important in education. AI-driven systems can analyze user data to tailor educational content to individual needs. For instance, PERSUADE, a system developed by Kaptein et al., uses machine learning algorithms to adapt persuasive messages to the user's preferences and behavior [36]. Similar approaches can be applied to language learning, where AI can adjust the difficulty level and content based on the learner's progress and preferences.

## 2.11 Chapter Conclusion

The reviewed literature demonstrates the potential of combining social chatbots, language models for poem generation, poetry's impact on self-confidence, and voice-based interfaces for language learning. However, despite the promising advancements in these areas, there is still a notable research gap. Existing applications and studies, while effective in isolated domains, often lack an integrated approach that combines these elements to specifically address the needs of adult learners in enhancing their self-confidence and verbal skills in a secondary language.

To the best of our knowledge, no current research or application has successfully merged social chatbots, poetic language models, and voice-based learning into a cohesive system aimed at improving both vocabulary and self-confidence through engaging and prolonged conversations. The Kavy project intends to fill this gap by developing a poetic voice companion that leverages AI and conversational technologies to provide adult learners with personalized, poetic interactions. This innovative approach not only aims to enhance vocabulary and conversational length but also to foster a deeper emotional and motivational connection through the unique use of poetry, thus addressing an unmet need in the field of language learning.

# CHAPTER 3

# RESEARCH METHOD

The development and evaluation of a mobile and personal voice companion named Kavy, specially designed to converse poetically with a human-like voice and capable of handling imperfect and immature user input, are introduced in this thesis. ChatGPT was employed for the back-end of Kavy, while the front-end comprises a React Native mobile app compatible with various platforms. Prompts were devised to generate poetic responses while also furnishing factual information in response to user inquiries.

## 3.1 Research Questions and Hypotheses

This study aims to explore the following research questions:

- *$RQ_1$: How can a voice companion support the language learning progress?*
- *$RQ_2$: What is the impact of a poetic version of the companion?*

To address these research questions, the following hypotheses were established:

- *[$H_1$]: The poetic voice companion will result in higher self-rated confidence compared to the non-poetic condition*
- *[$H_2$]: The participant will use new vocabulary when talking to a poetic companion*
- *[$H_3$]: The participant will have longer conversations with poetic voice companions compared to non-poetic*

Those people who learn a second language may find it difficult to use available assistants as their response depends on the user input accuracy. When a person makes any grammar mistakes in the speech, the assistant may not be able to interpret and provide the expected results. In this case, the user's self-confidence may be dropped preventing further engagement and leading to the end of the conversation.

## 3.2 Research Design

### 3.2.1 Overview

The research design employed for this study is a between-subjects design, focusing on comparing two versions of the voice companion: poetic and non-poetic. This design helps isolate the effects of the type of companion on the dependent variables, ensuring that any differences observed are due to the companion type and not other extraneous factors. A mobile app was developed as an instrument to conduct the user study.

### 3.2.2 Type of Study

This is a quantitative study, employing statistical methods to analyze numerical data collected from the participants. Quantitative methods are used to ensure objectivity and the ability to generalize findings to a larger population.

### 3.2.3 Variables

- Independent Variable: Type of voice companion (poetic vs. non-poetic). This variable is manipulated to observe its effect on the dependent variables.
- Dependent Variables: Self-rated confidence, new vocabulary usage, and conversation length. These variables are measured to assess the impact of the independent variable.

## 3.3 Materials and Methods

### 3.3.1 App Design

The Kavy app was developed using React Native for its cross-platform compatibility, allowing it to function seamlessly on both Android and iOS devices. The app features a minimalist interface designed to facilitate focused conversation. Key components of the app include:

- Voice Recognition: Utilizes built-in offline voice recognition services to accurately capture user speech.
- ChatGPT Integration: Employs the ChatGPT API for generating responses, ensuring high-quality natural language interaction.
- Google Neural2 Voice Models: Enhances conversation clarity with human-like voice synthesis for response audio playback.

### 3.3.2 Surveys and Questionnaires

The Foreign Language Self-Efficacy (FLSE) questionnaire proposed by Mills et al. [37] was adapted for this study (see Appendix 1 for the questionnaire). This well-validated survey measures language-speaking confidence levels. The same questionnaire was used for both pre-surveys and post-surveys to evaluate changes in self-confidence over the study period.

### 3.3.3 Data Logging Tools

All conversation data is collected from the frontend app and posted to the backend which is a Node Java Script (NodeJS) server hosted on the Google Cloud Run platform. The backend is responsible for analyzing the data and saving the analyzed information into MongoDB collections. Key data logging components include:

- Session Data: Logs the start and end times of user sessions.
- Conversation Data: Records each conversation's duration and content.
- Word Cloud: Tracks unique words spoken by users over time, including new words introduced in each conversation

## 3.4 Procedure

### 3.4.1 Study Phases

The study was conducted in several phases to ensure a structured and systematic approach:

1. Recruitment: Participants were recruited from university settings, targeting individuals interested in improving their English-speaking skills.
2. Pre-Survey: Participants completed a pre-survey to establish baseline measurements of self-confidence and language skills.
3. Daily Interaction: Over two weeks, participants were instructed to use the Kavy app daily for at least 5 minutes, engaging in conversations with the voice companion. All data was automatically logged.
4. Post-Survey: At the end of the study period, participants completed a post-survey to measure any changes in self-confidence and language skills.

### 3.4.2 Instructions to participants

Participants received detailed instructions on how to use the Kavy app, including how to initiate conversations and navigate the interface. They were encouraged to speak naturally and use the app regularly to maximize the study's effectiveness.

### 3.4.3 Ethical Considerations

Ethical approval was obtained from the NUS ethics committee before commencing the study. Participant consent was secured through the pre-survey, ensuring that they were fully informed about the study's purpose and procedures. Confidentiality and data privacy were strictly maintained throughout the study.

## 3.5 Data Collection

### 3.5.1 Methods

Data was collected through a combination of surveys and app interaction logs. Pre-surveys and post-surveys provided quantitative data on self-confidence and vocabulary usage. Interaction logs captured detailed information about each conversation, including duration and content.

### 3.5.2 Tools

The frontend app collected conversation data and posted it to a backend (NodeJS server hosted on Google Cloud Run). The backend processed and stored the data in MongoDB collections. Key metrics recorded included:

- Self-Rated Confidence: Measured through FLSE questionnaire responses.
- New Vocabulary Usage: Calculated based on the number of new words introduced in conversations.
- Conversation Length: Measured by the duration of each conversation.

## 3.6 Data Analysis Plan

### 3.6.1 Statistical Tests

T-Tests were used to analyze the data between subjects. These tests compared the means of the poetic and non-poetic groups to determine if there were statistically significant differences in self-rated confidence, new vocabulary usage, and conversation length.

### 3.6.2 Definition and Explanation of Metrics

The main metrics used in data analysis are derived from established linguistic measures and adapted to fit the context of the research. The primary metrics include the Average New Words Ratio and Speech Duration and Total Words. While Type-Token Ratio (TTR) and Lexical Density serve as supportive, validated measures, formulas are closely aligned with these established concepts to ensure rigor and relevance.

- **Average New Words Ratio**: Designed to evaluate the vocabulary acquisition of users interacting with Kavy. This metric calculates the ratio of new words spoken by each user to the total number of words spoken. It is inspired by the Type-Token Ratio (TTR), which assesses lexical variety by comparing the number of unique words (types) to the total number of words (tokens) in a text.

$$U_{\text{daily\_average\_newWords\_ratio}} = \overline{\left(\frac{|\text{userNewWords}|}{|\text{userAllWords}|}\right)_n}$$

– Where $n$ represents the day in the series.

– $|\text{userNewWords}|$ is the count of new words spoken by the user on a given day.

– $|\text{userAllWords}|$ is the total word count for that day.

– This approach is akin to the TTR, which measures lexical diversity and variety [38, 39].

- **Speech Duration and Total Words**: Evaluates user engagement by measuring the average number of words spoken per day and the duration of conversations. It draws on principles used in speech rate studies, where productivity is often measured in words per minute (wpm). Understanding the relationship between the number of words spoken and the duration of conversations helps gauge the depth and length of user interactions.

$$U_{\text{daily\_average\_allWords}} = \overline{|\text{userAllWords}|_n}$$

$$U_{\text{speech\_duration}} \propto |\text{userAllWords}|$$

– Where $|\text{userAllWords}|_n$ is the total word count for user $U$ on day $n$.

– Speech duration is proportionate to the total words spoken.

– This metric is supported by existing research on speech duration and word count correlations in language studies [40].

- **Type-Token Ratio (TTR)**: A widely used linguistic measure for assessing lexical diversity. It is calculated as the ratio of the number of unique words (types) to the total number of words (tokens). A higher TTR indicates greater lexical diversity, which is a desirable outcome in language learning.

$$\text{TTR} = \frac{\text{Number of Types}}{\text{Number of Tokens}} \times 100$$

– Where the number of types is the count of unique words.

– The number of tokens is the total word count.

– TTR is a validated measure that provides a benchmark for evaluating the effectiveness of vocabulary acquisition interventions [38, 39].

- **Lexical Density**: Measures the proportion of content words (nouns, verbs, adjectives, and adverbs) to the total number of words in a text. It provides insight into the complexity and richness of the language used. Higher lexical density indicates more complex language use, which can be an indicator of advanced language skills.

$$L_d = \frac{\text{Number of Lexical Items}}{\text{Total Number of Words}} \times 100$$

  - Where the number of lexical items includes all content words.
  - The total number of words encompasses both content and function words.
  - Lexical density is a well-established metric in computational linguistics and provides a solid foundation for assessing language complexity [40].

By aligning formulas with these well-established linguistic measures, it is ensured that the data analysis is both rigorous and meaningful. These metrics collectively help in evaluating the impact of Kavyconversational chatbot on user engagement, vocabulary acquisition, and overall language proficiency. This approach also addresses the feedback from the research supervisor, ensuring that the variables are meaningful and accurately represent the requirements of the study.

## 3.7 Limitations of the Study

### 3.7.1 Potential Biases

Potential biases in the study design or data collection methods include variations in participant engagement and the remote nature of the study, which might influence the results. Efforts were made to minimize biases through consistent instructions and regular reminders to participants.

### 3.7.2 Generalizability

The findings may have limited generalizability due to the small sample size and the specific demographic of university students in Sri Lanka. Future studies with larger and more diverse samples are needed to confirm the findings.

### 3.7.3 Technical Limitations

The study was highly dependent on an internet connection as the Kavy response is being generated in real-time via API calls. Both ChatGPT generation and Voice Synthesis require network calls. A slower connection in remote areas may affect the quality of the user experience.

## 3.8 Pilot Study

### 3.8.1 Purpose

A pilot study was conducted to identify any issues in the study design and to refine the data collection methods. The pilot aimed to test the functionality of the Kavy app and ensure that the data logging and analysis processes were effective.

### 3.8.2 Findings

The pilot study revealed that session time was not an accurate indicator of engagement due to potential idleness. Adjustments were made to save individual conversation times and improve word cloud analysis. Specifically, conversation durations were added to each conversation document, and new words were tracked per conversation. These changes led to faster and more accurate data processing.

## 3.9 Study Design and Methodology

### 3.9.1 Study Design

To investigate the research questions, two conditions were established: a poetic and non-poetic version of Kavy. A decision was made to proceed with a *between-subjects design* method due to significant learning effects. In the study, one independent variable is present: 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." Several dependent variables were measured 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.

### 3.9.2 User Group Selection

Users are automatically assigned to either a control or treatment group during their registration step. Kavy system back-end is designed to evenly spread the users among these two groups. To further enhance data manageability, virtual organizations were introduced, and treatment and control groups were pre-created per each organization. All these pre-created organizations were added to an MCQ question in the pre-survey so the user could select their organization. In the data analysis, Each cluster can be analyzed individually to understand their behavior. For instance, the University Students cluster may be different from other clusters.

To decrease noise and further validate the results, a series of questions were included to assess individuals' belief in their ability to use and navigate computer technologies effectively, known as *computer self-efficacy*. The question was adapted from a pre-validated questionnaire proposed by Compeau et al [41] (see Appendix 1 for the questionnaire). Measuring these factors will help mitigate the impact on the data in cases where individuals may face challenges in using modern mobile apps or computers.

#### 3.9.3 Measure FLSE - Both Surveys

Foreign Language Self-Efficacy (FLSE) questions mainly measure the language-speaking confidence level. This questionnaire was proposed by Mills et al. [37] (see Appendix 1 for the questionnaire). The same questionnaire was adapted to evaluate if someone's self-confidence can be improved by using Kavy. The same question set was included in both the pre-survey and post-survey, providing the capability to compare responses at the conclusion of the study.

In accordance with the literature [42], it may be important to perform a manipulation check to determine whether participants can perceive the difference between poetic and non-poetic responses. To address this, 10 brief audio responses were recorded from the poetic companion and 10 from the non-poetic version, each lasting no more than 10 seconds. These clips were included in the post-survey, where participants were asked to evaluate them using a 5-point Likert scale (Fig-5). The Qualtrics question was configured to randomize the order of the clips, mixing the poetic and non-poetic audio clips. Scoring was assigned based on the type of question: for non-poetic questions, a score of 5 was assigned to the non-poetic end of the Likert scale, while for poetic questions, a score of 5 was assigned to the poetic end of the Likert scale, with the opposite end receiving a score of 1.

#### 3.9.4 Participants

Participants were selected from a couple of universities in Sri Lanka, with an average age of 23 years ($M$=23.49, $SD$=3.68). All participants had adequate levels of computer literacy and self-efficacy, determined by the pre-survey responses ($M$=5.88, $SD$=1.27). Overall, 40 participants were recruited, of which 12 successfully completed the study, including all surveys.

#### 3.9.5 Procedure

The recruitment process commenced with the dissemination of meticulously crafted flyers, strategically placed across university campuses. Additionally, these flyers were distributed among friends and acquaintances attending various universities, who in turn

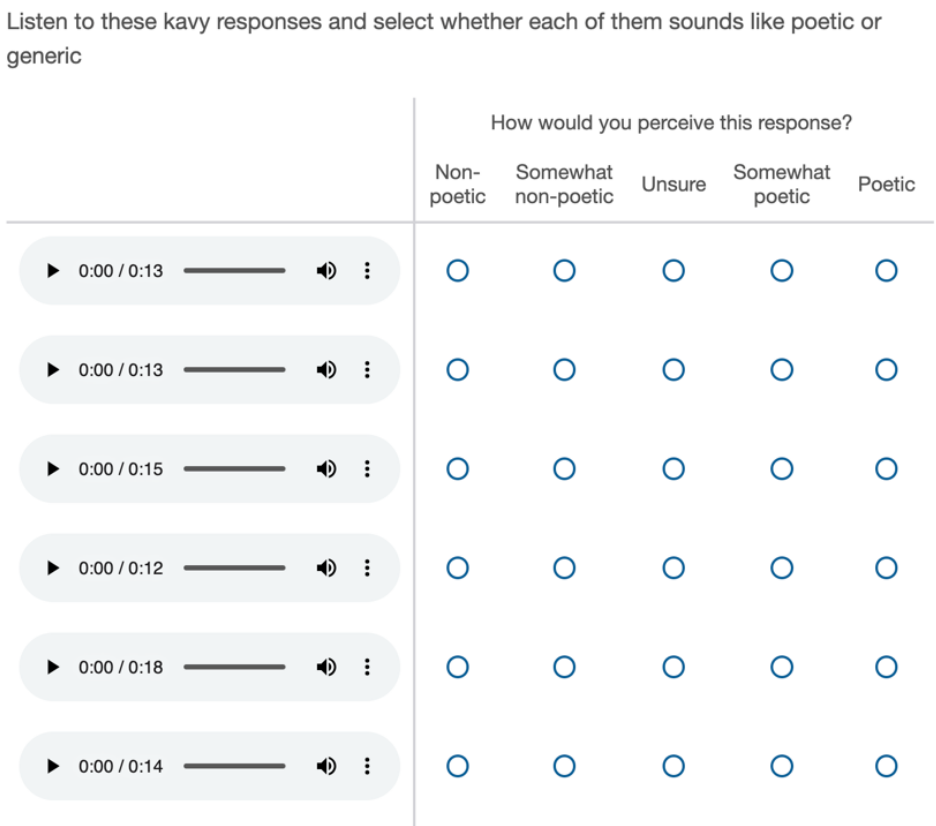



Fig. 3.1: Manipulation Audio Clips

shared them with potential participants. This grassroots approach ensured widespread visibility and accessibility to individuals who might be interested in participating.

These flyers not only provided a brief overview of the study but also contained Quick Response (QR) codes and download links, facilitating easy access to the Kavy app on both the App Store and Google Play Store. This multi-channel approach, combining physical distribution with personal referrals, maximized outreach and engagement with the target audience.

Upon downloading the app, participants experienced a seamless onboarding process without any cumbersome setup procedures. The login process, using Apple ID or Google ID, also served as a registration step, making participant enrollment quick and easy.

Upon successful login, participants were immediately directed to the pre-survey, strategically positioned on the app's home screen. This deliberate placement ensured maximum visibility and encouraged prompt completion, thereby expediting the enrollment process.

The pre-survey served as a crucial gateway to participation, providing researchers with invaluable demographic information and securing participant consent. Additionally, participants were prompted to select their respective organizations, laying the groundwork for data segmentation and analysis based on organizational affiliations.

Throughout the study duration, participants were tasked with engaging in daily interactions with the Kavy app, fostering a sense of familiarity and connection with the virtual companion. Daily push notifications served as gentle reminders, nudging participants towards consistent engagement and adherence to the study protocol.

Upon reaching the cumulative interaction threshold of 70 minutes (achieved over a period of 14 days with a minimum of 5 minutes per day), participants were automatically prompted to complete the post-survey. This survey, akin to its predecessor, served as a pivotal data collection instrument, marking the culmination of participants' involvement in the study.

## 3.10 Data Logging and Extraction

Self-reported question responses were collected through the pre and post-surveys to measure the language-speaking confidence level. This was based on Foreign Language Self-Efficacy (FLSE) questions [37] (see Appendix 1 for the questionnaire). In addition to this, throughout the study period, Kavy continuously reported all activities in real-time. These activities included logging session data and conversation data. Each session can have multiple conversations, and both conversation and session contain their duration in seconds.

Each conversation document is linked to a session by *'sessionId'*, and each session is linked to a user by *'userId'* data fields. All this data is securely saved in the database clusters. The session duration begins calculating as soon as the app opens and updates the database when the app is closed or sent to the background. Similarly, the conversation duration starts calculating once the user starts speaking and stops when Kavy stops speaking and saves the conversation record in the database.

When the user session is idle or in the background for more than 20 minutes, the app is designed to split the session by starting a new session and saving the previous one. In addition to these two data types, each user has extra documents. This also includes a *word cloud* document containing all the unique words spoken by the user and Kavy so far. Similarly, each conversation is associated with data arrays such as *all User Words, all Kavy Words, user new words, Kavy new words*, each specific to the conversation.

Day 1 marked the initiation of the study, with subsequent days being denoted as *relative dates* (Formula 3.4, Implementation 4.3) within a sequential time frame. The study was structured to operate remotely with minimal oversight. The utilization of relative dates facilitates the aggregation of user records, accommodating variations in the commencement of app usage among individual participants.

The *new word ratio* is calculated per conversation and aggregated these values as the sum for all users each day. The *new word ratio* is determined by dividing the number of new words by the total number of words. This metric offers insights into the usage of new words, as depicted in Formula 3.2. It served as the primary metric for assessing vocabulary usage. In addition to this primary metric, an extra analysis was conducted to examine word usage. This secondary measurement explained further below, considers how many new words are spoken per minute (Formula 3.3), providing a slightly different perspective. The corresponding chart can be found in Appendix 3. Related implementation details can be found in the next chapter (new words ratio implementation 4.2, new words per minute implementation 4.5)

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, the user's *conversation contribution* (Formula 3.1) value is calculated as a fraction (Implementation 4.4).

The *daily average new words per minute* value is calculated using the derived variable, assuming that the user's speech duration of a conversation is directly proportional to the user's total word count. The *daily average all words* (3.5) is then derived by counting the total words per conversation.

$$U_{\text{conversion\_contribution}} = \frac{|userAllWords|}{|userAllWords + keyAllWords|} \tag{3.1}$$

$$U_{\text{daily\_average\_newWords\_ratio}} = \overline{\left(\frac{|\text{userNewWords}|}{|\text{userAllWords}|}\right)_n} \tag{3.2}$$

$$U_{\text{daily\_average\_newWords\_per\_minute}} = \overline{\left(\frac{\left|\text{userNewWords} \times \left(\frac{60}{\text{duration}}\right) \times U_{\text{conversion\_contribution}}\right|}{\text{duration}}\right)_n} \tag{3.3}$$

$$sn = \text{relativeDay} = \left\lceil \frac{\text{conversation.createdAt} - \text{user.createdAt}}{1000 \times 60 \times 60 \times 24} \right\rceil \tag{3.4}$$

$$U_{\text{daily\_average\_allWords}} = \overline{|\text{userAllWords}|_n} \tag{3.5}$$

$$U_{\text{speech\_duration}} \propto |\text{userAllWords}| \tag{3.6}$$

$$K_{\text{speech\_duration}} \propto |\text{KavyAllWords}| \tag{3.7}$$

Where 'n' represents the day of the series [1-14]. 'U' represents "User", 'K' represents "Kavy"

## 3.11 Statistical Analysis Methods

In this study, pairwise T-Tests were used to analyze data between subjects in a design study. The T-Tests helped in comparing the means of two groups (poetic and non-poetic) to determine if there is a statistically significant difference between them.

### 3.11.1 Application of T-Tests

- Self-Confidence Improvement: A T-Test was applied to compare pre and post-survey FLSE scores between the two groups. The two-sample T-Test revealed a significant difference, indicating higher self-rated confidence in the non-poetic group compared to the poetic group.

- Vocabulary Improvement: A T-Test was used to compare the average new words ratio between the poetic and non-poetic groups. The results showed a significant difference, with the poetic group having a higher new word ratio.

- Conversation Length: A T-Test assessed differences in the average number of words spoken per day between the groups. The non-poetic group demonstrated longer conversation durations, indicating better engagement.

Assumptions such as the normality of distribution and equality of variances were checked for each T-Test. Any violations were addressed by applying appropriate statistical techniques to ensure the validity of the results.

## 3.12 Qualitative Feedback and User Experience

### 3.12.1 Qualitative Feedback

Participant feedback provided valuable insights into the user experience and areas for improvement. Below are some notable comments:

- Interface and Design: "App interface is too boring, it needs more graphical elements."

- Motivation: "They would like to see some statistics, real-time statistics on the screen which may motivate them."

- Complexity of Responses: "Poetic replies are too complex, that we need to simplify those responses."

- Engagement: "Gamification would help to improve more engagement."

### 3.12.2 User Experience Evaluation

To address the feedback, the following steps are proposed for future iterations of Kavy:

- Enhanced Interface: Incorporate more graphical elements to make the app visually appealing. This could include more vibrant colors, interactive elements, and visual feedback for user actions.
- Real-Time Statistics: Display real-time statistics on the screen to motivate users. For example, showing progress in vocabulary improvement, conversation length, and other metrics can encourage continuous usage.
- Simplified Responses: Simplify the poetic responses to ensure they are easy to understand. This may involve reducing the complexity of language and focusing on clear, concise communication.
- Gamification Elements: Add gamification features such as badges, leaderboards, and progress tracking to enhance engagement. These elements can make the learning process more enjoyable and competitive, encouraging users to interact with the app more frequently.

## 3.13 Chapter Conclusion

This chapter outlined the methodology used in the development and evaluation of Kavy, a mobile voice companion app designed to enhance language learning through poetic conversations. The app, built with NodeJS backend and a React Native frontend for cross-platform use, aimed to support users in improving vocabulary and self-confidence. The study design, data logging and extraction, statistical analysis methods, and qualitative feedback and user experience were all detailed in this chapter.

Two primary research questions were investigated: *how a voice companion supports language learning and the impact of a poetic rendition*. Three hypotheses guided the study: that the *poetic companion would boost self-confidence, increase new vocabulary usage, and lead to longer conversations compared to a non-poetic version*.

Kavy research employed a between-subjects design to compare poetic and non-poetic companions. Quantitative methods and T-Tests were used to analyze self-rated confidence, vocabulary usage, and conversation length. The app featured robust voice recognition, ChatGPT integration, and human-like voice synthesis.

Participants were recruited from a university setting and engaged with the app daily over two weeks. Data was collected through surveys and app interaction logs, and ethical considerations were addressed. The study revealed some limitations, including potential biases, limited generalizability, and initial technical challenges, which were mitigated through a pilot study.

In summary, this chapter provided a comprehensive methodology for evaluating Kavy, highlighting the app's potential impact on language learning and the benefits of poetic interactions.

# CHAPTER 4

# IMPLEMENTATION

An artifact, named Kavy, has been designed and developed. It comprises a mobile application featuring a minimalist front-end, tailored to facilitate focused conversation. The application is cross-platform compatible and functions optimally on both Android and iOS smartphones.

Given that modern mobile operating systems integrate built-in offline voice recognition services, leveraging these services proves highly efficient. Kavy utilizes native voice recognition services to ensure accuracy and optimal performance.

The entirety of the conversation is managed through the ChatGPT chat API [1], guaranteeing high-quality natural language interaction. To expedite response times, the frontend (the Kavy app) directly accesses the ChatGPT API without routing through the backend.

Google Neural2 voice models further enhance conversation clarity by providing human-like voice synthesis in the response audio playback. Synthesis API [2] is also called directly from the front end to prevent any potential network latency.

Accurate data logging is a crucial element of any research project, and the Kavy backend is specifically developed to manage all data logging requests. All conversation and session data are asynchronously posted to the backend in real-time. The backend subsequently saves these data logs into the respective collections within the database.

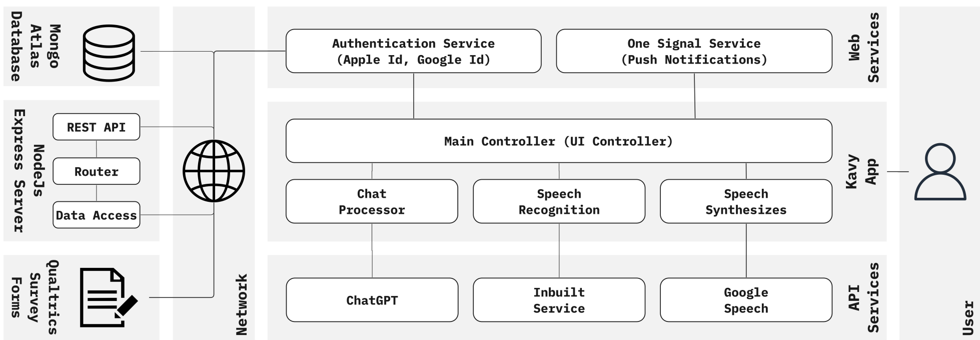



Fig. 4.1: Kavy's overall system diagram

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

[1]https://api.openai.com/v1/chat/completions
[2]https://texttospeech.googleapis.com/v1/text

## 4.1 System Design

The Kavy system comprises three major components, seamlessly connected services: a mobile front end, a back-end server, and a database. The front end was developed using the React Native cross-platform framework, enabling the creation of both Android and iOS apps from a single codebase. Communication between the front end and back end is handled by a NodeJS server instance, which connects to a MongoDB database. The overall architecture is illustrated in Figure 6.

## 4.2 Mobile app

The mobile app, developed using React Native, provides cross-platform support for both Android and iOS. The Google Text-to-Speech API was integrated for audio synthesis, and Kavy's voice was fine-tuned using Neural2 voice models to achieve a natural human-like quality. Native voice detection frameworks of the platforms were employed to recognize spoken user input. ChatGPT was utilized as the response generation service.

The user interface follows minimalist design principles (less is more [43], consistent design, visual clarity [44], simplicity [45]) to minimize user confusion and ensure ease of use. Users interact with the app through a simple mechanism: pressing and holding a button to speak, then releasing it to send the message. This interaction model, similar to the WhatsApp voice message button, is introduced early on the home screen for easy accessibility. The interface design is depicted in Figure 4.2.

## 4.3 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 JavaScript Object Notation (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 [46]. The prompts were designed in a way that resulted in a JSON object at the end. By performing some cleanups and validations, this object can be parsed by any software framework. Despite including the emotion field in the response structure, it's important to note that, within the current research context, this information is not being utilized.

Multiple prompt revisions were evaluated to achieve the optimal prompt in terms of response time and response quality. During development, the delay in web request-response was observed, and the optimal prompt that ensures a shorter waiting time was determined. An optimal prompt should also minimize decoding errors, which are considered a measure of quality. Below prompts are used in the Kavy app and the result takes a shorter waiting time and the quality is acceptable. Table 4.1 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"*

**TABLE** 4.1: SAMPLE USER INPUTS AND Kavy'S OUTPUTS FOR BOTH CONDITIONS

| **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? |

The pilot study participants reported a lack of accurate responses from the poetic version of Kavy. Feedback indicated that poetic responses often lacked substantial information or did not provide actual answers to questions. To address this, an extra phrase was added to the prompt: *"the first line of which should be an exact answer to the question in the input"*.

## 4.4 Data sanitization

GPT web responses are not always accurate in terms of expected response formatting. There is a chance they may contain some random noise and misformatting, as the response is generated by a language model. When the receiver does not handle this carefully, the system may crash, or the user experience could be compromised. To ensure the chat interface runs smoothly, a series of validations need to be performed before parsing the response to the main User Interface (UI) The following paragraphs explain different possibilities of errors and how they are handled.

Some GPT responses contained hidden ASCII control characters such as 'ˇ2029' (paragraph separator), 'ˇ0000' (null character), 'ˇ001F' (unit separator), 'ˇ007F' (delete), and 'ˇ009F' (application program command). These characters can cause issues because they are non-printable and often unexpected in text data. Whenever a response contains one of these characters, the application encounters a runtime compilation failure, as the default JSON deserialization method is unable to decode the body. A validation and cleanup method was added to scan those characters and replace them with an empty string.

Some GPT response message JSON bodies were incorrectly formatted with left quotation marks '“' and right quotation marks '”' instead of the straight quotation marks '"' that are required to properly wrap the JSON body. These curly quotation marks cause runtime failures during the decoding process because they are not recognized as valid JSON delimiters. To address this issue, a validation method was added to clean up these erroneous characters by replacing them with the expected straight quotation marks.

In some instances, GPT responses contain formatting errors related to JSON key formatting. Specifically, the key 'output' was sometimes enclosed in single quotation marks ('output':) or had no quotation marks at all (output:), rather than being properly enclosed in double quotation marks ("output":). Additionally, there were instances where the key 'Output' was capitalized instead of being in lowercase ('output'). These inconsistencies can cause issues during JSON parsing. To address these errors, the following corrections were implemented: replacing ('output':) with ("output":), replacing (output:) with ("output":), and correcting the case from 'Output' to 'output'. These changes ensure that the JSON keys conform to the correct formatting standards and prevent parsing errors

Well formatted response expected from the GPT service is as below.

Listing 4.1: Sample response from kavy

```
{
    "output": "Elon Musk, a visionary mind, birthed Tesla's
       drive, a car so divine. Electric whispers hum, as
       innovation sparks, revolutionizing roads, leaving
       fossil echoes in the dark.",
    "emotion": "happy",
    "question": "How has Elon Musk's vision impacted the
       automotive industry beyond Tesla?"
}
```

## 4.5 Survey development

Pre and post-survey forms were created and deployed on Qualtrics servers. The pre-survey form is designed to receive ‘userId‘ as an embedded data field (Figure 8) which can be passed as a Uniform Resource Locator (URL) parameter while opening the survey link. This ‘userId‘ is then used by a Qualtrics web service component (Figure 7) which is configured to make a web request to the Kavy back-end upon successful submission of the survey. This web call is captured by the Kavy back-end and updates the user's survey completion status on the 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 6, The back-end was developed 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. It was ensured that the user could not access the Kavy app without completing the pre-survey by blocking the UI and checking the survey completion status. This status is updated at the end of the Qualtrics survey submission using a Qualtrics web service and calling Kavy back-end method. 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. It was decided to leave the app-enabled even after the study period.

## 4.6 Security and Privacy

Ensuring the security and privacy of user data is crucial for Kavy. Several measures have been implemented to protect user information and comply with data protection regulations.

### 4.6.1 Data Security Measures

All data is stored securely in a MongoDB cluster. The survey forms managed by National University of Singapore (NUS) Qualtrics services are also configured to maximize security, ensuring that user responses are protected. User registration and login are securely managed by Google Authentication or Apple Authentication, and these login details are not transported to or saved in the backend.

### 4.6.2 Anonymization and Privacy

To protect user privacy, direct identity information is not attached to data logs. Conversations are not exposed in any public locations. The analytics page only displays calculated statistics without revealing individual identities or conversation details. This approach ensures that user data remains confidential and is used solely for research purposes.

### 4.6.3 Compliance with Regulations

Kavy adheres to relevant data protection regulations, including the General Data Protection Regulation (UI) These regulations require strict controls over personal data, and compliance is automatically enforced by the App Store and Play Store during the app publishing process. Kavy has implemented necessary measures to ensure compliance, such as obtaining user consent for data collection and providing users with the ability to request data deletion.

## 4.7 Performance Evaluation

To ensure that Kavy delivers a seamless user experience, a comprehensive performance evaluation was conducted. This evaluation focused on optimizing response times and resource usage.

### 4.7.1 Direct API Access

The frontend directly accesses the ChatGPT Application Programming Interface (API) and Google Text-to-Speech API to minimize latency.

### 4.7.2 Dockerized Backend

The backend was dockerized and deployed on Google Cloud Run, utilizing a serverless architecture that automatically scales resources based on demand. This setup ensures that the app can handle a high volume of concurrent requests.

### 4.7.3 Resource Management

The backend is configured to handle up to 50 instances simultaneously, ensuring that data logging and other backend processes run smoothly without overloading the system.

## 4.8 User Experience Metrics

Evaluating the user experience (UX) is crucial for understanding how users interact with Kavy and the overall effectiveness of the application. Several UX metrics were identified and measured to assess the app's usability, engagement, and satisfaction. These metrics were evaluated using the User Experience Questionnaire (UEQ) [3].

The key UX metrics used in this study include:

- **Supportiveness**: Measures how much the app supports the user in completing their tasks. A supportive app helps users achieve their goals smoothly.
- **Ease of Use**: Assesses how simple and straightforward it is to use the app. A high ease of use indicates that users find the app uncomplicated and user-friendly.
- **Efficiency**: Evaluates how quickly and effectively users can complete tasks using the app. An efficient app allows users to achieve their goals with minimal effort.
- **Clarity**: Measures how clear and understandable the app is to users. High clarity means that users can easily comprehend and navigate the app.
- **Excitement**: Assesses how engaging and stimulating the app is to users. An exciting app captures users' interest and motivates continued use.
- **Interest**: Evaluates how interesting the app is to users. An interesting app keeps users engaged and encourages them to explore more features.
- **Inventiveness**: Measures the app's innovation and creativity. High inventiveness indicates that the app offers unique features and a fresh experience.

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

- **Leading Edge**: Assesses how modern and cutting-edge the app feels to users. A leading-edge app provides a sense of being at the forefront of technology.

These UX metrics were collected through the UEQ administered in the post-survey. The responses were analyzed to determine the overall user experience of Kavy and to compare the experiences between the poetic and non-poetic voice companions.

## 4.9 Technical Challenges and Solutions

The development of Kavy presented several technical challenges, particularly in ensuring consistent performance across different platforms and managing data effectively. This section outlines these challenges and the solutions implemented to address them.

### 4.9.1 Cross-Platform Development

React Native was chosen for its cross-platform capabilities, allowing the app to be developed for both Android and iOS using a single codebase. However, some libraries did not perform consistently across both platforms, requiring platform-specific code changes to maintain the same behavior. In cases where a library was not supported on one platform, alternative libraries with cross-platform support were selected.

### 4.9.2 Data Management

Managing data effectively was crucial for the app's performance and user experience. The backend was dockerized and deployed on Google Cloud Run, leveraging a serverless architecture to automatically scale resources based on demand. This setup ensured that the backend could handle a high volume of concurrent requests, primarily for data logging.

### 4.9.3 Security and Privacy

Ensuring the security and privacy of user data was another significant challenge. All data was stored in a secure MongoDB cluster, with survey forms managed by NUS Qualtrics services for maximum security. Data logs were anonymized, and conversations were not exposed publicly, ensuring that user privacy was maintained.

## 4.10 Data Extraction

All the recorded data is piped to Mongo charts dashboards, which provide a clear understanding at a glance and also offer ways to export data for any further analysis. Mongo data aggregation queries were written to achieve this, as described in Mongo

Query Language (MQL) . Derived data analysis formulas, as discussed in the previous chapter (See formulas 3.1), are implemented using the Mongo descriptive queries shown below.

Listing 4.2: New Words Ratio

```
{
  "$divide": [
    { "$size": "$userNewWords" },
    { "$size": "$userAllWords" }
  ]
}
```

Listing 4.3: Relative Date

```
{
  "$ceil": {
  "$divide": [
    {
      "$subtract": ["$createdAt", "$user.createdAt"]
    },
    86400000 /* Number of milliseconds in a day (1000 ms * 60
       s * 60 min * 24 hours) */
  ]
}
}
```

Listing 4.4: User Words fraction

```
{
  "$divide": [
    { "$size": "$userAllWords" },
    {
      "$add": [
        { "$size": "$userAllWords" },
        { "$size": "$kavyAllWords" }
      ]
    }
  ]
}
```

Listing 4.5: User's new words per minute

```
{
  "$multiply": [
```

```
    60,
    {
      "$multiply": [
        { "$size": "$userNewWords" },
        {
          "$multiply": [
            { "$divide": [1, "$duration"] },
            { "$divide": [1, "$userWordsFraction"] }
          ]
        }
      ]
    }
  ]
}
```

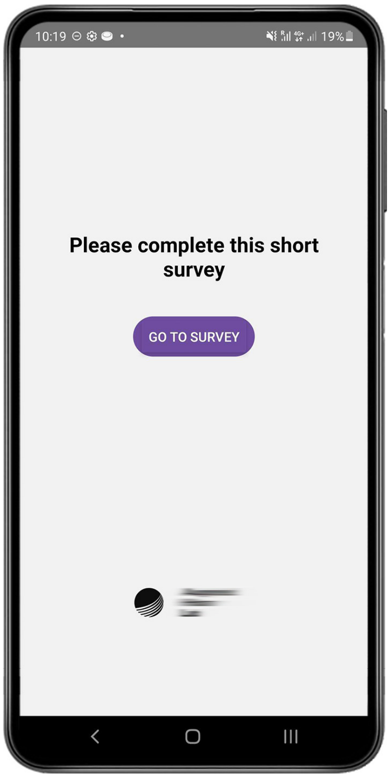
## 4.11 Chapter Conclusion

This chapter outlined the implementation of Kavy, a cross-platform mobile application designed to facilitate language learning through poetic conversations. The app leverages modern mobile operating systems' built-in offline voice recognition services, the ChatGPT chat API for managing conversations, and Google Neural2 voice models for human-like audio responses. The minimalist interface and direct API access ensure a smooth user experience.

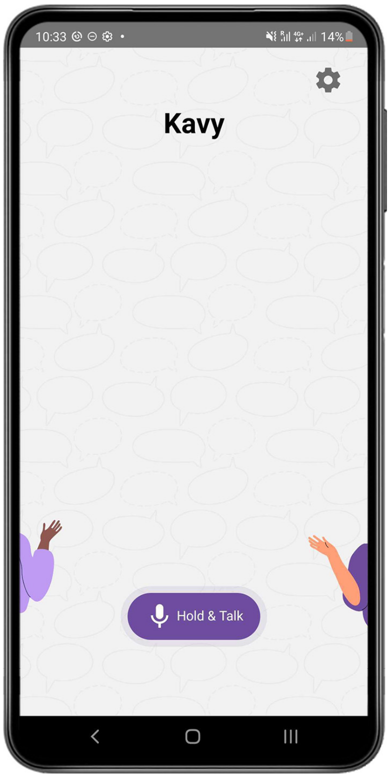
The system architecture integrates a React Native frontend with a NodeJS backend and MongoDB database, enabling seamless data logging and analysis. Prompt engineering was a critical aspect, optimizing response quality and speed. Extensive data sanitization ensures reliable performance, while survey development and integration with Qualtrics provide robust data collection mechanisms.

Security and privacy are prioritized through data anonymization, adhering to regulations like GDPR. Performance evaluations and optimization strategies ensure the app functions efficiently across various devices. Addressing technical challenges, such as cross-platform consistency and data management, further enhanced the app's reliability.

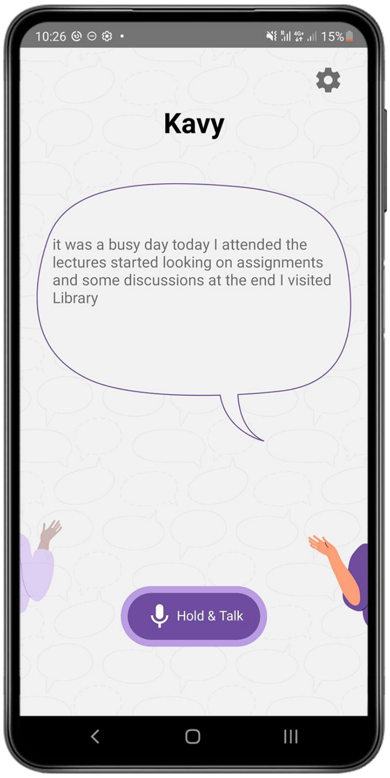
In summary, this chapter provides a comprehensive overview of the technical implementation of Kavy, highlighting the integration of advanced technologies and meticulous development processes that contribute to its effectiveness as a language learning tool.




(a) Survey prompt




(b) Home screen




(c) User speaks


(d) Kavy speaks

Fig. 4.2: 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.

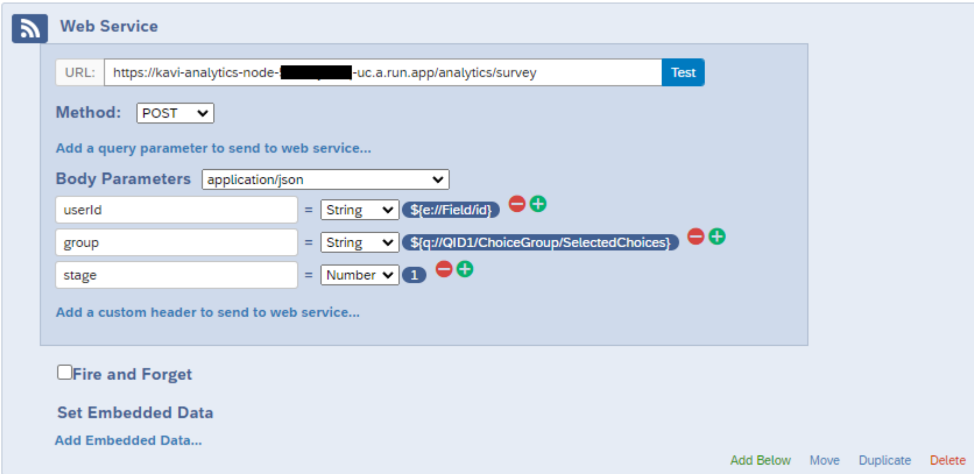



Fig. 4.3: Qualtrics Web Service Configuration

*Note.* The Qualtrics web service is configured to send a POST message to the Kavy backend with the following body parameters: userId, assigned group, and survey stage (0 for pre-survey and 1 for post-survey)

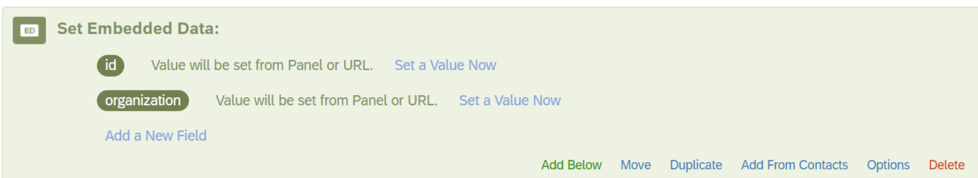



Fig. 4.4: Qualtrics Embedded Data

*Note.* User's Id and Organization (Id) are configured to be collected from URL parameters as embedded data in Qualtrics survey forms

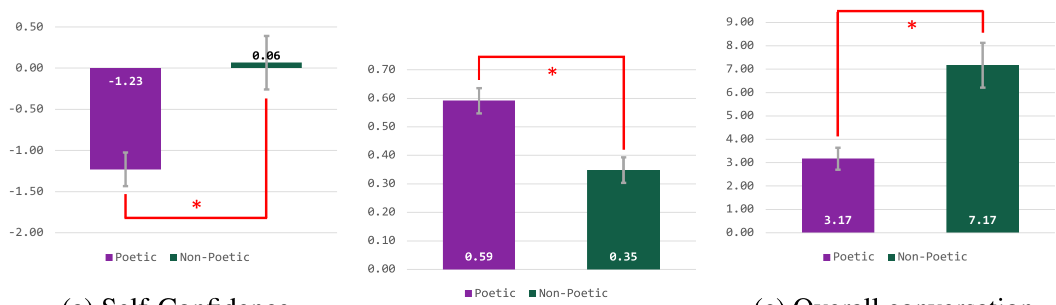
# CHAPTER 5

# RESULTS AND ANALYSIS

This chapter presents the results and analysis of the study conducted using Kavy. The analysis focuses on participant demographics, improvements in self-confidence and vocabulary, the ability to foster longer conversations, patterns of adaptive learning, user experience, study design, and data logging. Additionally, it discusses the statistical methods used, qualitative feedback, challenges and limitations, and future research directions.

## 5.1 Participant Demographics

At the end of the study period, 14 participants (6 from one group and 8 from the other) completed all tasks. Consequently, 2 participants were randomly excluded to ensure balanced groups. These 12 participants were included in the statistical analysis. The participants were recruited from a couple of universities in Sri Lanka, with an average age of 23 years (*M*=23.49, *SD*=3.68). All participants reported having adequate levels of computer literacy and self-efficacy, as indicated by their pre-survey responses (*M*=5.88, *SD*=1.27). The survey responses also indicated that, on average, these participants had been learning or practicing English for almost 14 years (*M*=13.93, *SD*=4.62). Participants showed 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. The question was adapted from a questionnaire proposed by Hamby et al. [47]. (see Appendix 1 for the questionnaire).




(a) Self-Confidence Difference Pre vs. Post

(b) Overall new words ratio average

(c) Overall conversation length average

Fig. 5.1: T-Test Results

## 5.2 Improvement in Self Confidence

Pre and post-survey responses were analyzed to measure changes in self-confidence, specifically focusing on the FLSE score. Table 5.1 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 $T(13) = -4.77$ revealed a strong ($p < 0.05$) significant difference (see Figure 5.1a), indicating that the non-poetic voice companion results in higher self-rated confidence compared to the poetic condition. Therefore *H1* is rejected.

### 5.2.1 Discussion on Self-Confidence Findings

The analysis revealed that the non-poetic voice companion resulted in higher self-rated confidence compared to the poetic condition. This finding contradicts the initial hypothesis (H1), which posited that a poetic voice companion would increase self-confidence. Several factors might explain this unexpected outcome.

Firstly, the poetic nature of responses may have introduced complexity and ambiguity, leading to confusion rather than confidence. Existing literature suggests that clear and straightforward communication is crucial in language learning contexts, especially for beginners [19, 20]. The complexity of poetic language might have overwhelmed the participants, reducing their perceived self-confidence.

Secondly, the familiarity and simplicity of non-poetic responses could have made users feel more at ease, thereby boosting their confidence. This aligns with findings from Törnqvist's study, which indicated that clear, non-poetic communication positively impacts language learners' self-confidence [20].

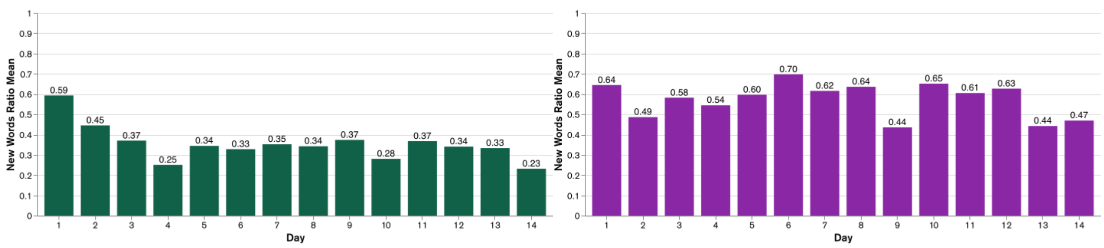
Lastly, the non-poetic group might have benefited from more direct and factual feedback, which is essential for building confidence in language learning [24]. The poetic group's feedback, while creatively engaging, might not have provided the straightforward reinforcement needed to enhance self-confidence.

## 5.3 Improvement in Vocabulary

A derived formula, namely *average new words ratio* (Equation 3.2), was analyzed. This formula calculates the ratio of the number of new words spoken by each user to the total number of words spoken by that user which is then Aggregated across all users within the group. This ratio is represented for each day as shown in Figure 5.2. A pairwise *t*-Test $T(12) = 7.17$ showed a significant difference ($p < 0.05$). The poetic group showed a higher new word ratio than the non-poetic group (see Figure 5.1b). Therefore, *H2* is accepted, indicating that users with the poetic version picked up more vocabulary.

| **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** 5.1: Foreign Language Self Efficacy pre and post values




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

(b) Daily average new words ratio: poetic group

Fig. 5.2: Daily average new words ratio analysis (14 days span)

## 5.4 Fostering Longer Conversations

The *average total words* (Equation 3.5) was calculated 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, as shown in Figure 5.3. 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 (see Figure 5.1c). Therefore *H3* is rejected indicating that a non-poetic voice companion creates longer engagement.

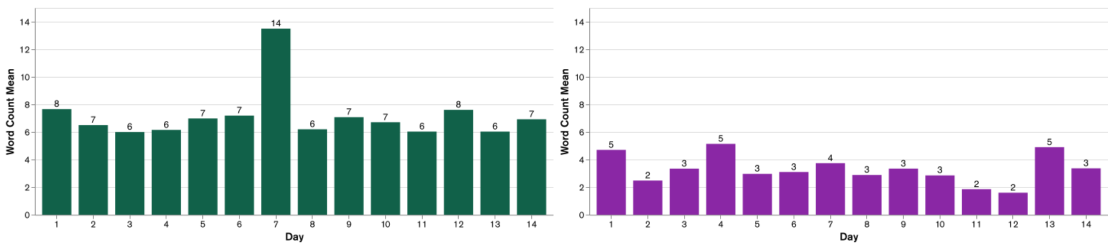
### 5.4.1 Discussion on Conversation Length Findings

The results indicated that the non-poetic voice companion led to longer conversations compared to the poetic version, rejecting our hypothesis (H3) that poetic companions would engage users in longer conversations. This outcome can be analyzed through several lenses.

One possible explanation is that poetic responses, while engaging, may not sustain prolonged interaction due to their complexity. Participants may have found it challenging to continuously respond to and interpret poetic language, leading to shorter conversations. This aligns with the findings of Saad et al., who noted that the complexity of poetry could sometimes hinder rather than help engagement in certain contexts [21, 22].

Furthermore, the non-poetic companion's straightforward responses likely facilitated easier and more fluid conversations. This simplicity could have made interactions feel more natural and less cognitively demanding, allowing participants to engage for longer periods. Research by Hijazi et al. supports this notion, highlighting that clear and simple communication often leads to more sustained interactions in language learning [24].

Another factor could be the participants' familiarity with non-poetic language in everyday communication. The poetic format, though novel and interesting, may have introduced an additional layer of difficulty that hindered sustained engagement. Finch's exploration of poetry in language learning suggests that while poetry can enhance creativity, it may not always be the best medium for sustained conversational practice [25].




(a) Daily total word count - control group (b) Daily total word count - treatment group

Fig. 5.3: Daily word count analysis (14 days span)

## 5.5 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 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. This question was adapted from a questionnaire proposed by Eric et al. [48] (see Appendix 1 for the questionnaire)

## 5.6 User Experience

The UEQ measures[1] were evaluated for both groups. The intention behind this was to understand how users would experience a poetic chat partner and whether a poetic condition has an impact on the overall UX. The key UX metrics assessed included attractiveness, perspicuity, efficiency, dependability, stimulation, and novelty. 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), as shown in Figure 10. Both groups indicated positive values for pragmatic quality and hedonic quality.

Both groups equally reported that Kavy is supportive regardless of whether 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 to be comparatively interesting (See Figure 9 for more insights).

The detailed analysis of these UX metrics provides valuable insights into how the design and interaction style of Kavy can be optimized to improve user satisfaction and engagement.

## 5.7 Intrinsic Movivation

Apart from the main measurements, a few additional factors were assessed in the post-survey. The data collected from both the treatment and control groups revealed insights into intrinsic motivation across different dimensions. For the treatment group, *intrinsic motivation towards accomplishment* [49] showed an average satisfaction level of 5.67

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

for improving English literacy weaknesses and skills, slightly higher than the control group's 5.50 and 5.67, respectively. Similarly, *intrinsic motivation to know* [49] indicated that the treatment group enjoyed learning and using the Kavy app, with means of 5.33 and 5.67, while the control group's means were 5.50 and 5.17. Interestingly, the control group reported higher enjoyment in the *intrinsic motivation to experience* [49] the stimulation dimension, with means of 5.50 and 5.67 compared to the treatment group's 5.00 and 4.67. These findings suggest that while the poetic voice companion was effective in enhancing certain aspects of intrinsic motivation, further investigation is needed to understand the variability in user experience enjoyment between the groups. Details of the questionnaire are provided in Appendix 2.

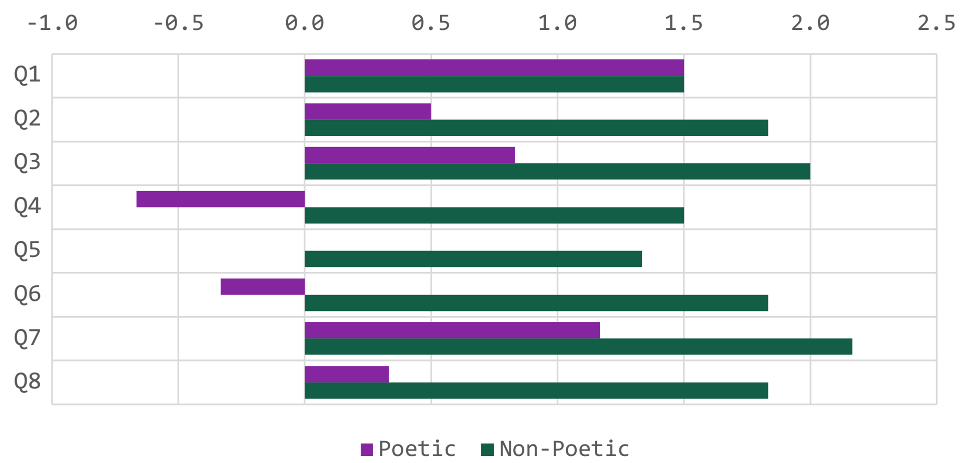



Fig. 5.4: User Experience Details

*Note.* 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

Both groups equally reported that Kavy is supportive regardless of whether 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 to be comparatively interesting.

## 5.8 Challenges and Limitations

### 5.8.1 Challenges

The study faced several challenges, primarily due to its remote nature:

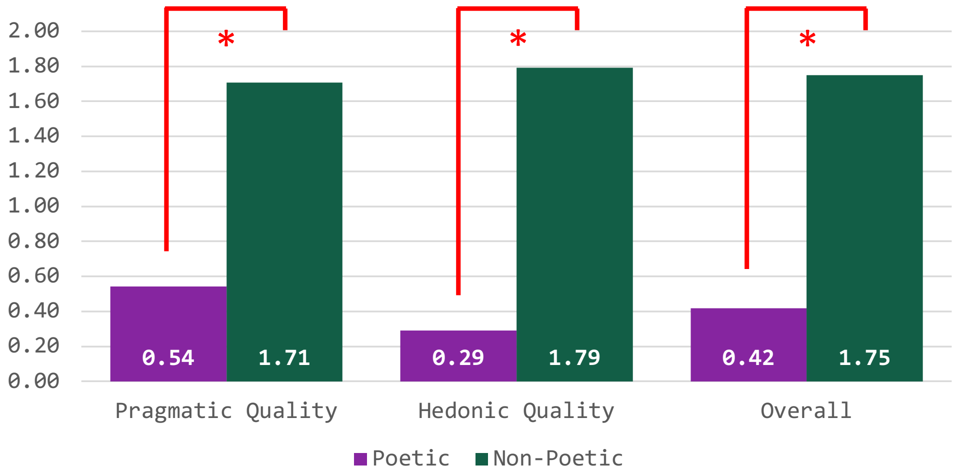



Fig. 5.5: User Experience Summary

- Participant Engagement: Ensuring consistent app usage among participants was difficult. Some participants did not use the app daily, took breaks, or dropped out after a few days. Despite push notifications and direct reminders from coordinators, maintaining daily engagement was challenging.

- Communication: The only way to communicate with participants was through push notifications or direct reminders via a coordinator. This mode of communication was sometimes insufficient to ensure consistent engagement.

### 5.8.2 Limitations

- Data Integrity: Due to inconsistent usage and dropouts, some data points had to be removed, potentially affecting the accuracy of the results. The removal of these data points, although necessary to maintain the quality of the research, could lower the overall accuracy and reliability of the findings.

- Sample Size: The final sample size of 12 participants may limit the generalizability of the findings. A larger sample size would provide more robust data and improve the reliability of the results.

- Remote Study Constraints: The remote setup introduced variability in how participants engaged with the app, which could influence the study outcomes. Differences in environment, distractions, and varying levels of commitment could have impacted the results.

## 5.9 Future Research Directions

### 5.9.1 Future Research Directions

Based on the findings and challenges of this study, several areas for future research are suggested:

- Larger Sample Size: Conducting the study with a larger sample size to improve the reliability and generalizability of the results. A more diverse participant pool can also provide insights into different user experiences and needs.

- Controlled Environment: Implementing the study in a more controlled environment to ensure consistent usage and minimize dropouts. A controlled setting can help in monitoring participant engagement more effectively and provide more accurate data.

- Enhanced Features: Exploring the impact of adding graphical elements, real-time statistics, simplified responses, and gamification on user engagement and learning outcomes. Future iterations of Kavy could include these enhancements to test their effectiveness.

- Longitudinal Studies: Extending the study duration to observe the long-term effects of using Kavy on language learning and user engagement. Longitudinal studies can provide deeper insights into how the app impacts language skills over time.

- Qualitative Research: Conducting qualitative research to gain deeper insights into user experiences and preferences. Interviews, focus groups, and open-ended survey questions can provide rich data that complements

### 5.9.2 Implications for Future Research

The findings and subsequent rejections of hypotheses H1 and H3 provide valuable insights and directions for future research. Understanding the limitations and challenges posed by poetic companions is crucial for refining language learning tools.

Future studies should explore ways to balance the creative engagement of poetic responses with the clarity and simplicity needed for effective language learning. This might involve developing adaptive response systems that adjust the complexity of responses based on the user's proficiency and comfort level.

Additionally, incorporating elements that enhance user confidence, such as more personalized feedback and interactive support mechanisms, could help mitigate the issues observed with poetic companions. Longitudinal studies that track user progress over extended periods might provide deeper insights into the long-term impacts of different conversational styles on language learning.

Overall, the rejection of these hypotheses underscores the need for a nuanced approach to integrating creative elements into language learning tools, ensuring they support rather than hinder learner engagement and confidence.

## 5.10 Chapter Conclusion

This chapter presented the results and analysis of the study conducted using Kavy. The analysis covered participant demographics, improvements in self-confidence and vocabulary, the ability to foster longer conversations, and adaptive learning patterns.

The study involved 12 participants from universities in Sri Lanka, with an average age of 23 years. The data indicated that the non-poetic voice companion led to higher self-rated confidence, while the poetic companion resulted in a higher new word ratio. However, the non-poetic group engaged in longer conversations, suggesting better overall engagement.

Patterns of adaptive learning showed participants' dedication to expanding their vocabulary and maintaining positive social perceptions. The user experience evaluation revealed that the non-poetic companion was preferred, being perceived as clearer and more interesting. Qualitative feedback highlighted areas for improvement, such as adding graphical elements, real-time statistics, simplifying responses, and incorporating gamification.

The study faced challenges, including maintaining consistent participant engagement and dealing with the limitations of a remote setup. Future research directions suggest larger sample sizes, controlled environments, enhanced app features, longitudinal studies, and qualitative research to gain deeper insights into user experiences and preferences.

In summary, this chapter provided a comprehensive analysis of the study's findings, highlighting the impact of poetic and non-poetic voice companions on language learning and user engagement. The results offer valuable insights for further development and research on enhancing language learning tools.

# CHAPTER 6
# CONCLUSION AND FUTURE WORK

## 6.1 Conclusion

In this research, the effectiveness of using a poetic voice companion was studied with the primary aim of improving adult learners' second language speaking skills in terms of self-confidence and vocabulary. The user study conducted showed that users' vocabulary could be improved when using poetic voice companions. It was presumed that users would engage in longer conversations with poetic companions. However, it was found that this was not the case, and longer conversations were had with non-poetic voice companions. Self-confidence was measured via a self-reported survey, though no significant results were observed. Furthermore, findings indicated that any voice companion would help improve adult learners' language-speaking skills. Notably, poetic chatbot users used a significant number of new words compared to non-poetic versions, suggesting potential long-term vocabulary improvement. These preliminary findings are based on a limited participant pool and a short study, indicating short-term effects or improvements of Kavy. Further studies are necessary to strengthen these findings and determine the long-term benefits of Kavy.

## 6.2 Limitations

### 6.2.1 Scope Limitations

The scope of this study is limited to adult learners in university contexts in Sri Lanka. Participants were all in the STEM stream, where the poetic nature is believed to be unfamiliar, and their first language is Sinhala. The findings may not be generalizable to other populations or settings without further research.

### 6.2.2 Methodological Limitations

Potential methodological limitations include variations in participant engagement due to the remote nature of the study and the small sample size, which may affect the reliability and generalizability of the results.

## 6.3 Future Work

### 6.3.1 Technical Enhancements

Certain challenges were underlined by the study, limiting the generalizability of the findings. Some of these limitations are of a technical nature. Currently, Kavy can only remember a few of the last conversations, and responses are made accordingly. Additionally, the user's demographics are not considered by Kavy, resulting in a lack of personalized chat experiences. In the future, it is envisioned that all collected user details, such as data from the pre-survey, will be fed into the initial GPT prompt. Furthermore, the prompt may be informed by other variables generated dynamically based on user activity such as screen time, current location, and physical activity. In this manner, Kavy may be transformed into a versatile companion, resembling an actual person.

### 6.3.2 User Engagement and Motivation

Another challenge was retaining the interest and motivation of the participants during the study period. In the main study group, out of 40 registered participants, only 12 completed the study. Many participants dropped out during the study period, and some engaged less frequently than expected. The 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.

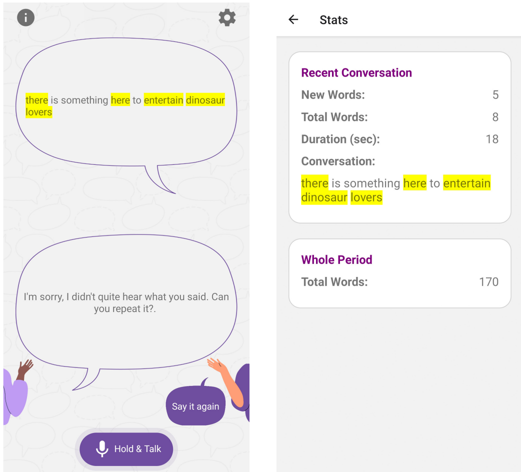
For future studies, it is crucial to incentivize users to interact consistently over the entire study period. To achieve this, incorporating gaming elements such as badges and ranks could be beneficial. Gamification can significantly enhance user engagement and motivation by providing tangible rewards for consistent usage.[50–53].

### 6.3.3 Realtime Progress Statistics

User engagement can be further enhanced by displaying real-time statistics of vocabulary improvements on screen (Figure 11.b). The initial version of this feature has already been tested and deployed with the latest version of the mobile app. This statistics screen includes the *Total Words Count* for the entire period, *New Words Count*, *Duration*, and the *Conversation Text* of the most recent conversation. Although the main user study did not benefit from this feature, future studies will take advantage of it. All distinct words spoken by the user are saved in individual word clouds. An additional network call is sent in the middle of each conversation as soon as the user finishes speaking. This allows the frontend to retrieve the new words from the user's conversation text promptly. As a result, the app highlights all new words even before Kavy finishes the conversation, providing the user with a real-time statistical presentation that may motivate them to speak more.

### 6.3.4 Generalizability and Participant Diversity

Participants were only recruited from Sri Lanka. Future studies should aim to recruit more participants from diverse ethnic, cultural, and religious backgrounds. This ap-




(a) (b)

Fig. 6.1: Kavy Latest Features

proach will help elicit more design choices and enhance the robustness of the study. By including a more diverse participant pool, the findings can be generalized to a broader population.

### 6.3.5 Interface and Interaction Improvements

The previous chat interface (which was used for the main study) consists of a basic chat bubble UI displaying the user's speech and Kavy's reply. The most recent chat interface (11.a) has improved to overcome this drawback by showing both chat bubbles on the same screen. It also includes a button that the user can press to listen again to the Kavy's reply which may help to improve listening skills. Engagement could be improved by incorporating a realistic avatar that animates as the companion speaks. Adapting existing humanoid digital face avatar plugins and services for this purpose is feasible and could make the interaction more engaging and lifelike.

### 6.3.6 Dynamic Learning Experience

The learning experience can be dynamically adapted to users' skill levels by analyzing the complexity of conversations. With the system supporting dynamic prompting, the mobile app can adjust prompts to elicit optimal responses that align with users' current language skills. Utilizing LLMs for analyzing grammar and conversation complexity is feasible. This dynamic adaptation can be further enhanced by incorporating mobile phone sensor data. Given that modern devices are equipped with location sensors and physical activity sensors (such as pedometers and accelerometers), Kavy's responses can be tailored based on the user's current context. Context-aware responses have the potential to significantly motivate users to engage in conversations, which will be explored in future studies.

### 6.3.7 Emotion Analysis and Empathy

Valuable insights into both past and future conversations can be provided by emotion analysis. Through the analysis of previous interactions, a better understanding of users' emotional states and preferences can be gained, enabling the tailoring of Kavy's vocal features to be more empathetic and dynamic. This involves adjusting parameters such as pitch, speaking rate, and speaking style to better align with users' emotional needs and enhance the overall conversational experience.

### 6.3.8 Collaborative Learning and Gamification

To further enhance the learning experience, incorporating elements of collaborative learning and gamification can be beneficial. Implementing leaderboards and collaborative challenges can foster a sense of community among users, encouraging them to engage more frequently and learn from each other. Gamification elements such as points, badges, and levels can provide additional motivation and make the learning process more enjoyable.

## 6.4 Research Extensibility

The primary focus of Kavy's research is language studies and poetic expression. Modifying the LLM prompts can easily adapt this focus to other domains. The research can also be enhanced by gathering and analyzing input queries across various prompts to produce distinct outputs, tailored to each domain. This adaptability makes the Kavy system a valuable tool for analyzing knowledge and skills in different fields. By embedding domain-specific knowledge within prompts, the system can respond accurately to each spoken input, generating relevant responses within the specified domain.

## 6.5 Reusability of artifacts and workflow adaptability for future studies

A number of artifacts as research deliverables have been produced during this research, which can be used for similar studies. The primary software components, including the frontend and backend repositories, are available on GitHub and can be shared upon request from researchers. Additionally, MongoDB snapshots and Qualtrics survey templates are available to accompany the main repositories. Researchers are welcome to contact us (either the lead researcher or supervisors) to request access, fork our pipeline, and receive guidance on implementation.

## 6.6 Final Thoughts

The introduction of Kavy, a poetic voice companion, has shown promise in improving vocabulary and engagement in language learning. However, several challenges and limitations were identified that need to be addressed in future research. By incorporating technical enhancements, improving user engagement strategies, and expanding participant diversity, the effectiveness of Kavy can be further validated. The potential for a dynamic, context-aware, and empathetic voice companion tailored to individual learners' needs holds significant promise for the future of language learning technology. Further research is required to explore these possibilities and establish long-term benefits.

# REFERENCES


[1] T. Hayrettin, "The relationship between self-confidence and learning turkish as a foreign language," *Educational research and reviews*, vol. 10, no. 18, pp. 2575–2589, 2015.

[2] A. M. De Groot and J. G. Van Hell, "The learning of foreign language vocabulary," *Handbook of bilingualism: Psycholinguistic approaches*, pp. 9–29, 2005.

[3] A. H. Kuppens, "Incidental foreign language acquisition from media exposure," *Learning, media and technology*, vol. 35, no. 1, pp. 65–85, 2010.

[4] S. Cooray, C. Hettiarachchi, V. Nanayakkara, D. Matthies, Y. Samaradivakara, and S. Nanayakkara, "Kavy: Fostering language speaking skills and self-confidence through conversational ai," in *Proceedings of the Augmented Humans International Conference 2024*, 2024, pp. 226–236.

[5] S. Valtolina and L. Hu, "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*, 2021, pp. 1–5.

[6] F. Braun, L. Block, and S. Stegmüller, "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, 2021, pp. 436–442.

[7] T. Kaluarachchi, C. Sankha, E. Wen, V. Tang, T. M. H. Nguyen, P. Chua, and S. Nanayakkara, "Kavi: A poet in your pocket that listens to your thoughts and whatnot!" 2022.

[8] L. Zhou, J. Gao, D. Li, and H.-Y. Shum, "The design and implementation of xiaoice, an empathetic social chatbot," *Computational Linguistics*, vol. 46, no. 1, pp. 53–93, 2020.

[9] H. Chin, H. Song, G. Baek, M. Shin, C. Jung, M. Cha, J. Choi, and C. Cha, "The potential of chatbots for emotional support and promoting mental well-being in different cultures: Mixed methods study," *Journal of Medical Internet Research*, vol. 25, p. e51712, 2023.

[10] R. Vesselinov and J. Grego, "Duolingo effectiveness study," *City University of New York, USA*, vol. 28, no. 1-25, 2012.

[11] L. D'Souza and D. Mimno, "The chatbot and the canon: Poetry memorization in llms," *Proceedings http://ceur-ws. org ISSN*, vol. 1613, p. 0073, 2023.

[12] A. Dai, "Gpt-2 for emily dickinson poetry generation," 2021.

[13] P. Sawicki, M. Grzes, F. Goes, D. Brown, and M. Peeperkorn, "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)."

[14] A. Kirmani, "Artificial intelligence-enabled science poetry, acs energy letters, 8, 574-576," 2022.

[15] M. Ghazvininejad, "Neural creative language generation," Ph.D. dissertation, University of Southern California, 2018.

[16] J. H. Lau, T. Cohn, T. Baldwin, J. Brooke, and A. Hammond, "Deep-speare: A joint neural model of poetic language, meter and rhyme," *arXiv preprint arXiv:1807.03491*, 2018.

[17] Z. Wang, W. He, H. Wu, H. Wu, W. Li, H. Wang, and E. Chen, "Chinese poetry generation with planning based neural network," *arXiv preprint arXiv:1610.09889*, 2016.

[18] J. Zhang, Y. Feng, D. Wang, Y. Wang, A. Abel, S. Zhang, and A. Zhang, "Flexible and creative chinese poetry generation using neural memory," *arXiv preprint arXiv:1705.03773*, 2017.

[19] C. Cronin and C. Hawthorne, "'poetry in motion'a place in the classroom: Using poetry to develop writing confidence and reflective skills," *Nurse education today*, vol. 76, pp. 73–77, 2019.

[20] S. Törnqvist, "Speaking through poetry-using spoken word poetry to lower speaking anxiety among swedish efl learners," 2019.

[21] S. Saad, R. Amat, R. Nordin, A. Hasan, and A. N. F. Zolkifli, "Poetry out loud: fostering esl learners public speaking skill," 2023.

[22] ——, "From poetry out loud to thinking out loud: The esl students' experience," *Social and Management Research Journal (SMRJ)*, vol. 20, no. 2, pp. 153–164, 2023.

[23] G. E. Chandler, "A creative writing program to enhance self-esteem and self-efficacy in adolescents," *Journal of Child and Adolescent Psychiatric Nursing*, vol. 12, no. 2, pp. 70–78, 1999.

[24] S. Mustafa, E. Melonashi, F. Shkembi, K. Besimi, and N. Fanaj, "Anxiety and self-esteem among university students: comparison between albania and kosovo," *Procedia-Social and Behavioral Sciences*, vol. 205, pp. 189–194, 2015.

[25] A. Finch, "Using poems to teach english. english language teaching. 15 (2), 29–45." *English Language Teaching*, vol. 15, no. 2, pp. 29–45, 2003.

[26] K. Simecek and K. Rumbold, "The uses of poetry," pp. 309–313, 2016.

[27] S. Bibauw, T. François, and P. Desmet, "Conversational agents for language learning: state of the art and avenues for research on task-based agents," in *CALICO, Date: 2015/05/27-2015/05/29, Location: Boulder, CO*. CALICO, 2015.

[28] S. Ruan, L. Jiang, Q. Xu, Z. Liu, G. M. Davis, E. Brunskill, and J. A. Landay, "Englishbot: An ai-powered conversational system for second language learning," in *26th international conference on intelligent user interfaces*, 2021, pp. 434–444.

[29] H.-H. Huang, A. Cerekovic, K. Tarasenko, V. Levacic, G. Zoric, I. S. Pandzic, Y. Nakano, and T. Nishida, "Integrating embodied conversational agent components with a generic framework," *Multiagent and grid systems*, vol. 4, no. 4, pp. 371–386, 2008.

[30] E. Katsarou, F. Wild, A.-M. Sougari, and P. Chatzipanagiotou, "A systematic review of voice-based intelligent virtual agents in efl education," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 18, no. 10, pp. 65–85, 2023.

[31] J. Petrovic and M. Jovanovic, "Conversational agents for learning foreign languages–a survey," *arXiv preprint arXiv:2011.07901*, 2020.

[32] G. Dizon, "Using intelligent personal assistants for second language learning: A case study of alexa," *Tesol Journal*, vol. 8, no. 4, pp. 811–830, 2017.

[33] A. C. Graesser, S. Lu, G. T. Jackson, H. H. Mitchell, M. Ventura, A. Olney, and M. M. Louwerse, "Autotutor: A tutor with dialogue in natural language," *Behavior Research Methods, Instruments, & Computers*, vol. 36, pp. 180–192, 2004.

[34] D. McDuff and M. Czerwinski, "Designing emotionally sentient agents emotionally sentient systems will enable computers to perform complex tasks more effectively, making better decisions and offering more productive services."

[35] A. Maley, "Overview: Creativity–the what, the why and the how," *Creativity in the English language classroom*, pp. 6–13, 2015.

[36] M. Kaptein, P. Markopoulos, B. De Ruyter, and E. Aarts, "Personalizing persuasive technologies: Explicit and implicit personalization using persuasion profiles," *International Journal of Human-Computer Studies*, vol. 77, pp. 38–51, 2015.

[37] N. Mills, F. Pajares, and C. Herron, "A reevaluation of the role of anxiety: Self-efficacy, anxiety, and their relation to reading and listening proficiency," *Foreign language annals*, vol. 39, no. 2, pp. 276–295, 2006.

[38] https://www.sltinfo.com/wp-content/uploads/2014/01/type-token-ratio.pdf#:~:text=URL:.

[39] https://www.sltinfo.com/type-token-ratio/.

[40] https://en.wikipedia.org/wiki/Lexical_density.

[41] D. R. Compeau and C. A. Higgins, "Computer self-efficacy: Development of a measure and initial test," *MIS quarterly*, pp. 189–211, 1995.

[42] D. J. Hauser, P. C. Ellsworth, and R. Gonzalez, "Are manipulation checks necessary?" *Frontiers in psychology*, vol. 9, p. 998, 2018.

[43] G. Hart, "Ten design principles and when to violate them."

[44] S. Krug *et al.*, "Don't make me think, revisited," *A Common Sense Approach to Web and Mobile Usability*, 2014.

[45] D. Norman, *The design of everyday things: Revised and expanded edition*. Basic books, 2013.

[46] Y. Wu and G. Hu, "Exploring prompt engineering with gpt language models for document-level machine translation: Insights and findings," in *Proceedings of the Eighth Conference on Machine Translation*, 2023, pp. 166–169.

[47] A. Hamby and D. Brinberg, "Happily ever after: How ending valence influences narrative persuasion in cautionary stories," *Journal of Advertising*, vol. 45, no. 4, pp. 498–508, 2016.

[48] E. M. Anderman, T. Urdan, and R. Roeser, "The patterns of adaptive learning survey," *What do children need to flourish? Conceptualizing and measuring indicators of positive development*, pp. 223–235, 2005.

[49] X. Li, J. P.-A. Hsieh, and A. Rai, "Motivational differences across post-acceptance information system usage behaviors: An investigation in the business intelligence systems context," *Information systems research*, vol. 24, no. 3, pp. 659–682, 2013.

[50] J. P. Gee, “What video games have to teach us about learning and literacy,” *Computers in entertainment (CIE)*, vol. 1, no. 1, pp. 20–20, 2003.

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

[52] T. W. Malone and M. R. Lepper, “Making learning fun: A taxonomy of intrinsic motivations for learning,” in *Aptitude, learning, and instruction*. Routledge, 2021, pp. 223–254.

[53] M. Prensky, “Digital game-based learning,” *Computers in Entertainment (CIE)*, vol. 1, no. 1, pp. 21–21, 2003.

# APPENDIX A

# PRE SURVEY QUESTIONAIRE

### Pre-study

### Attitude toward literacy skills

1. Use a Likert scale of 1 – 7 for all the following items: (1) strongly disagree, (4) neutral, (7) strongly agree.
2. 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.
3. 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

1. Use a Likert scale of 1 – 7 for all the following items: (1) strongly disagree, (4) neutral, (7) strongly agree.
2. 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.
3. 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

1. Please rate the following scales to answer the following statements (1–7): (1) no chance, (4) average, and (7) completely certain.

2. 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.

3. 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.

**Patterns of Adaptive Learning Survey (PALS)**

1. Use a Likert scale of 1 – 7 for all the following items: (1) strongly disagree, (4) neutral, (7) strongly agree.

2. 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 the 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 like I have trouble with my English.

3. Source: Eric M. Anderman, Tim Urdan, and Robert Roeser (2006) The patterns of adaptive learning survey, Chapter 14 (Why do children need to flourish).

**Demographic Information**

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

# APPENDIX B

# POST SURVEY QUESTIONAIRE

### Post-study

### Intrinsic motivation toward accomplishment

1. Use a Likert scale of 1 – 7 for all the following items: (1) strongly disagree, (4) neutral, (7) strongly agree.
2. 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.
3. 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

1. Use a Likert scale of 1 – 7 for all the following items: (1) strongly disagree, (4) neutral, (7) strongly agree.
2. 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.
3. 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**

1. Use a Likert scale of 1 – 7 for all the following items: (1) strongly disagree, (4) neutral, (7) strongly agree.

2. 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.

3. 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**

1. Use a Likert scale of 1 – 7 for all the following items: (1) strongly disagree, (4) neutral, (7) strongly agree.

2. 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.

3. 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**

1. Based on your use of the Kavy app, please rate the following scales to answer the following statements (1–7): (1) no chance, (4) average, and (7) completely certain.

2. 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.

3. 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.

# APPENDIX C

# ADDITIONAL FIGURES AND CHARTS