languages

MDPI

Article

ChakapBot: A Generative AI-Powered Chatbot for the Revitalisation of Baba Malay

Nala H. Lee [1,*] and Huiyu Zhang [2]

[1] Department of English, Linguistics & Theatre Studies, Faculty of Arts & Social Sciences, National University of Singapore, Singapore 117570, Singapore
[2] School of Informatics & IT, Temasek Polytechnic, Singapore 529757, Singapore; zhang_huiyu@tp.edu.sg
* Correspondence: nala.lee@nus.edu.sg

Abstract

This paper reports on the pilot implementation of ChakapBot, a generative AI-supported chatbot developed to support both the revitalisation and documentation of Baba Malay, an endangered heritage language in Singapore. A six-week pilot was conducted with 26 participants, the majority of whom were adult learners with limited prior exposure to the language and disrupted intergenerational transmission. Drawing on surveys, the study examines participant engagement, perceived usefulness, and the role of the chatbot within a blended learning model combining in-person instruction and digital practice. ChakapBot was trained on a community-curated corpus comprising a dictionary, a textbook, and a grammatical description of Baba Malay, positioning it as both a learning support tool and a means of operationalising existing documentation resources for community use. Findings indicate that participants valued the chatbot for its flexibility, accessibility, and low-pressure environment for independent practice, and perceived it as a complement to human-led instruction rather than a replacement. Participants also reported using Baba Malay beyond classroom settings, including in the home, workplace, and community events. The study highlights how ethically designed, community-curated AI tools can bridge language documentation and revitalisation by transforming documented linguistic knowledge into accessible, everyday practice for learners.

Keywords: language documentation; language revitalisation; endangered languages; Baba Malay; generative AI; chatbots; blended learning; community-led design

Academic Editors: K. David Harrison and Lisa Lim

Received: 9 February 2026
Revised: 29 June 2026
Accepted: 3 July 2026
Published: 8 July 2026

1. Introduction

1.1. Contextualising ChakapBot for Baba Malay Revitalisation

This article examines the use and effectiveness of artificial intelligence (AI), in particular GenAI (generative AI) for language instruction, both in conjunction with and in contrast to in-person language classes, as a means of revitalising Baba Malay in Singapore. Specifically, ChakapBot was trained on a community-curated corpus comprising a textbook, a dictionary, and a grammatical description of Baba Malay developed through a language documentation project. This article therefore explores an innovative approach to bridging documentation and revitalisation with AI, as well as examines the outcomes of such an approach. While the role of AI has been discussed to some extent, with regard to the preservation and revitalisation of endangered languages (Ingram, 2025; Wang, 2024), most of these discussions broadly address the effectiveness of these potential programmes from a general or computational perspective, but do not address the actual outcomes of

actual revitalization initiatives, particularly from the perspective of participants themselves. In fact, when it comes to language revitalisation programmes in general, scholars have observed that there is very little research on their outcomes, despite the existence of what are plausibly hundreds or even thousands of revitalisation projects worldwide (O'Grady & Hattori, 2016, p. 53). Against this backdrop, the present article evaluates the effectiveness of a GenAI chatbot in language revitalisation that leverages existing documentation and language documentation resources alongside in-person instruction, as assessed through participants' perspectives.

Baba Malay is a critically endangered contact language spoken primarily in Malaysia and Singapore (A. G. L. Lau, 1984; Lee, 2019). It developed through early intermarriages between Southern Min–speaking Chinese traders and Indigenous Malay-speaking women, with historical records indicating that Chinese voyagers had settled in the Malay Archipelago as early as the 15th century (Lee, 2022). The descendants of these intermarriages are known as the Peranakan Chinese, and their home language came to be known as Baba Malay. Baba Malay was historically spoken mainly in Malacca, in present-day Malaysia (Tan, 1979). However, with the British settlement of Singapore in the early 19th century, many Peranakan Chinese migrated there for trade and commerce. Today, it is estimated that there are fewer than 1000 speakers of Baba Malay in Malacca, just as there most likely are fewer than 1000 speakers in Singapore (Lee, 2022). As the project is based in Singapore, it is useful to provide an overview of the sociolinguistic context there:

Early adoption of an English-medium education and changing demographic trends account for the minimal intergenerational transmission of Baba Malay (Lee, 2019). The Peranakan Chinese were among the first in the region to adopt an English-medium education, following the establishment of English-medium schools by the British (Ansaldo et al., 2007). The Peranakan Chinese, who benefited from their knowledge of English by serving as successful intermediaries between the British administrators and the newly arrived Chinese immigrants, began introducing English into their homes, thereby replacing Baba Malay. Another key factor contributing to Baba Malay's endangerment relates to demographic changes. Today, intermarriages between Chinese and Malay speakers no longer result in a Peranakan Chinese identity. While arranged marriages between Peranakan Chinese families were once common—or Chinese men would marry into Peranakan Chinese households—such practices are no longer the norm. Most Peranakan Chinese now marry outside their community, which in turn affects their choice of home language (Lee, 2019).

Baba Malay has never gained official status in Singapore. Today, it is spoken alongside the official languages of English, Malay, Mandarin, and Tamil. Among these, English serves as the primary medium of government, commerce, and education; Malay is the national language; and Mandarin is taught as a subject to Chinese students (Rappa & Wee, 2006). Administratively, Peranakan Chinese are classified as Chinese and are required to learn Mandarin at school as a subject. Other non-official languages spoken in Singapore include Hokkien, Teochew, Cantonese, and Hakka—alongside various Indo-Aryan, Dravidian, and Austronesian languages (Jain, 2021). As a point of comparison, approximately 157,000 people reported using Hokkien most frequently at home (Department of Statistics Singapore, 2020), making it the largest non-official language in Singapore, whereas fluent Baba Malay speakers number fewer than 1000.

Efforts at revitalising Baba Malay have been uneven, even though the question of Baba Malay's vitality was raised as early as the 1980s (Chia, 1983; A. G. L. Lau, 1984; Pakir, 1986). While the language experienced a modest revival through its usage in *wayang Peranakan* 'Peranakan plays' when these regained popularity (Peranakan theatre was popular pre-World War II), the scope of this revival remained somewhat limited in domain. The publication of an annotated gloss (Gwee, 1993) and dictionary (Gwee, 2006) followed in

subsequent decades. The language has also been the focus of a language documentation project resulting in a grammar (Lee, 2014, 2022), but it was only in 2017 that Baba Malay language classes were first organised by the Gunong Sayang Association, a cultural association that focuses on the promotion of Peranakan culture in Singapore. The language classes, *Mari Kita Chakap Baba* 'Let us speak Baba' were subsequently led by community member, Baba Kenneth Chan, who published an accompanying textbook, *Mari Chakap Baba* 'Let's speak Baba' (Chan, 2018), that incorporated dialogues recorded by Peranakan actors from the Gunong Sayang Association. However, interest in these classes waned after the COVID-19 pandemic.

Notably, one of the main issues with previous iterations of Baba Malay classes was the lack of interest from younger generation. In 2019, when the class was still running, there were nine learners in their 20s and 30s, and almost twice as many older participants, with 17 learners aged 40 and above. This was, in fact, the term that enrolled the highest number of participants under the age of 40. Baba Malay language classes also have had a high level of attrition. While the 2019 class started out with 26 learners, the last lesson was attended by14 learners after 8 weeks. In addition to dealing with the challenge of sustaining the interest of learners, the crux of the issue for language revitalisation remains the general lack of engagement among younger learners.

Successful language revitalisation is contingent upon there being intergenerational transmission. As O'Grady notes, "the most telling measure of language vitality is intergenerational transmission" (O'Grady & Hattori, 2016, p. 45). Lee and Van Way (2018) highlights that there is a fundamental assumption that "intergenerational transmission is necessary for the continued viability of any language" and that this assumption underlies major frameworks for assessing language vitality, including UNESCO's nine factors for assessing language vitality (UNESCO Ad Hoc Expert Group on Endangered Languages, 2003); the Expanded Graded Intergenerational Disruption Scale (EGIDS) (M. P. Lewis & Simons, 2010) utilised on Ethnologue (www.ethnologue.com); and the Language Endangerment Index (LEI) utilised by the Catalogue of Endangered Languages (ELCat), which is hosted on the Endangered Languages Project web portal (www.endangeredlanguages.com) (also see O'Grady & Hattori, 2016). If a language is not transmitted to younger generations, it will eventually cease to exist. It is noteworthy that O'Grady and Hattori (2016, p. 45) go as far as to say that

> "Intergenerational transmission, the ultimate goal of language revitalization efforts, can only be achieved by (re)establishing the conditions under which an imperiled language can be acquired by the community's children."

Although establishing or reestablishing those conditions in a sustainable way is a tall order for the Peranakan Chinese community for all the reasons that Baba Malay is threatened, the notion that the youth should be targeted must be taken seriously. While it is valuable to engage the broader community in language revitalisation projects, it also becomes problematic when participants are predominantly from older generations. The target audience of revitalisation initiatives should ideally include children, or at least individuals of child-bearing age where possible, as they are the ones who might one day transmit the language to their own children.

Given the challenges that the Peranakan Chinese community has faced in engaging the youth, the current project was designed to encourage participation among younger learners. The inclusion of the AI chatbot component, capable of effectively and efficiently operationalising existing documentation materials, was intended to appeal to younger learners and foster greater engagement among them than previous revitalisation efforts. This study also contributes to emerging work on AI-assisted language revitalisation by demonstrating how a retrieval-constrained GenAI system can operationalise existing lan-

guage documentation while simultaneously generating new, validated linguistic data within its development workflow.

1.2. AI Projects for Language Revitalisation

The origins of modern AI can be traced to the 1950s, when Alan Turing posed the seminal question, "Can machines think?" (Turing, 1950). However, recent years have witnessed an exponential acceleration in AI advancement. Following the emergence of voice-recognition personal assistants such as Siri and Alexa in the 2010s, the 2020s have ushered in a new technological era marked by the proliferation of GenAI models, including OpenAI's ChatGPT and Google Gemini. It is this latter form of AI that is most evoked by the public these days, although a diverse range of AI technologies exist. Subsequently, a variety of language revitalisation initiatives have emerged that employ artificial intelligence through diverse methodologies and applications.

In India, Adi Vaani is conceived as an app that translates between Hindi and smaller languages—it aims to promote these smaller languages, while also helping speakers access online information and government services (Dasgupta, 2025). In Japan, Tatsuya Kawahara and team have used AI in the form of speech recognition to automate the process of transcribing 300 to 400 h of oral Ainu data, and is now being used to generate speech from text, so that people in Hokkaido can learn the Ainu language (J. Lau, 2025). Tools such as AI Pirika which combine chatbot and speech recognition technology have also been developed for Ainu (Blanks, n.d.). In New Zealand, Te Hiku Media, a non-profit media organisation, has utilised an AI model for speech recognition and language processing. Te Hiku Media primarily focuses on Māori language revitalisation—with AI, it has produced a resource for automatic bilingual transcriptions and the Rongo app for practising Māori pronunciation, among others (Te Hiku Media, n.d.). Based in Canada, the First Languages AI Reality (FLAIR) initiative, headed by Michael Running Wolf has a foundational automatic speech recognition system that is aimed at the rapid creation of custom models for endangered languages—these can be used for language learning, audio transcription, and voice-controlled technology among other uses (Mila—Quebec Artificial Intelligence Institute, n.d.).

Recent innovative projects emerging from North America exemplify the growing use of artificial intelligence in language revitalisation but also question the space for GenAI. Notable examples include *Masheli*, a chatbot developed by Jacqueline Brixey, and *Skobot*, created by Danielle Boyer. *Masheli* draws upon a corpus of traditional animal stories to facilitate communication in both Choctaw and English (Brixey & Traum, 2020), while *Skobot* is an Anishinaabe-speaking robot designed to be worn on the shoulder, resembling a small parrot (The Steam Connection, 2024). The latter operates on an internally developed AI model capable of learning and reproducing language (Jampel, 2025). Both technologies provide interactive, conversational opportunities that contribute to sustaining and revitalising Indigenous languages. Notably, neither of these systems utilise GenAI in what appears to be ethnical, conscious, and cautious decisions to preserve accuracy, community ownership, and cultural sensitivity. As Brixey states, "We should have a right to say how our languages are used" and that "ChatGPT could be good in Choctaw, but it's currently ungrammatical…it makes up what it claims are tribal stories" (Jampel, 2025).

It is clear that the impact of AI-generated content can be overwhelmingly negative when the AI generated content is erroneous, plagiarised or misleading: A series of Abenaki books that were being sold online on Amazon were found to have incorrect translations and made up words, and were falsely attributed to a real Indigenous language expert, in addition to authors that could not be found online (Glorieux-Stryckman, 2024). Inadvertently, GenAI has been avoided or approached much more cautiously than not in language revital-

isation. That is not to say the GenAI has been completely avoided. In Iceland, Þorsteinsson and team have been fine-tuning and training GPT-4 to generate Icelandic content, using Reinforcement Learning from Human Feedback—which allows testers to provide a prompt, then select the best outcome, and further edit the outcome so that responses can be more informed in the future (OpenAI Press Release, 2025).

Against this backdrop of risk and caution, the role of community engagement becomes especially salient. Even without AI, community engagement itself is an important component of successful language revitalisation programmes (Wiltshire et al., 2024). The use of AI makes it even more imperative that community is engaged at every step of the way. In a chapter that discusses the potential role of AI in language preservation and revitalisation, based on interviews with speakers of minority/endangered languages, Ingram (2025, p. 62) states that "native speakers play a critical role in ensuring that AI tools reflect the true character of their language" and that "[b]y engaging communities in these efforts, AI developers can create tools that are culturally sensitive, linguistically accurate, and tailored to the needs of the language's speakers." Community engagement can not only empower the speakers whom revitalisation is meant for, but also "ensure they retain ownership and authority" (Ingram, 2025, p. 69). Relatedly, Indigenous communities have long expressed concerns that existing international intellectual property mechanisms offer insufficient protection for Indigenous knowledge and cultural expressions, particularly in relation to over-commercialisation, misuse, and digital appropriation (UNESCO, 2022). This has in the past been the source of conflict between community and external organisations, such as in the Lakota Language Consortium affair (Brewer, 2022). It becomes more imperative than ever that community engagement is mandatory in any knowledge enterprise that is AI generated or otherwise.

Last but not least, it is useful to consider the notion of framing of AI "as a complement to, rather than a replacement for, traditional methods of cultural and linguistic preservation." (Ingram, 2025, p. 73). In that vein, this project that the article describes does not aim to replace the traditional teacher in the classroom, in the context of language revitalisation, but shows the extent of how AI can be utilised in combination with traditional in-person classes, and examines what it is not able to do, *vis-à-vis* the traditional classroom approach.

Despite the many cautions that must be heeded, there remain compelling reasons to explore how GenAI might be harnessed in the context of language revitalisation. AI has the benefits of "scalability" and "efficiency" (Ingram, 2025, p. 73), and as Soroush Vosughi states in an interview, "generative AI and large language models significantly lower barriers to revitalizing endangered languages, rapidly producing valuable linguistic resources even from minimal data" (Barath, 2025). AI therefore becomes compelling when framed as an innovative means of linking existing language documentation with revitalisation efforts.

In the case of Baba Malay, beyond engaging younger generations through a novel approach, GenAI is also envisaged as offering speakers a scalable, modular, and low-cost model for language revitalisation, given the availability of a small but carefully curated body of documentation.

2. Materials and Methods

2.1. The Core of ChakapBot: Team and Technology

ChakapBot (chakap means 'to speak' in Baba Malay) was developed through a community-led collaboration involving Peranakan Chinese members and academic partners. The project brought together the first author, a community member-linguist; Baba Malay educator Baba Kenneth Chan; and heritage expert Baba Andy Gwee, working in close collaboration with a core development team led by Dr Zhang Huiyu at Temasek Polytechnic, Singapore. Community member Nyonya Josephine Tan was also part of

the core development team. To further support community engagement, the project was backed by the Peranakan Association Singapore, with financial support provided through a grant from Singapore's National Heritage Board.

ChakapBot is an AI-powered chatbot developed for instructional use in Baba Malay. It draws on a set of community-curated resources, including A Baba Malay Dictionary, compiled by the late Baba William Gwee (Gwee, 2006), Baba Chan's Baba Malay textbook, *Mari Chakap Baba* (Chan, 2018) and a grammatical description of the language developed part of the first author's language documentation project (Lee, 2014). Permissions were sought and granted from the authors, who are all members of Peranakan Chinese community, and from Baba William Gwee's son, Baba Andy Gwee—it was imperative that contributors make an informed decision about consenting to the use of their materials to train a GenAI chatbot. Note that all of these materials focused on Baba Malay as spoken in Singapore, as the study's locus is Singapore.

The chatbot was developed as a modular system. The system comprises three interrelated components: a learner-facing chatbot, a web-based application for recording and validating audio, and an internal AI Training Hub used for resource management and governance. Together, these components form an integrated digital environment that supports structured language learning, spoken language documentation, and the controlled use of generative AI. The design of each component, and the ways in which they interact, are described in detail below:

ChakapBot incorporates a hybrid instructional chatbot that combines intent-driven, rule-based interactions with retrieval-constrained GenAI (Basar et al., 2023; Long & Zhang, 2024), with each component serving a distinct pedagogical role. Instead of operating as a fully open-ended chatbot, the system differentiates between structured language-learning tasks that require accuracy and consistency, and more exploratory language queries that allow for limited conversational flexibility.

At the core of ChakapBot's instructional design is Google Cloud Platform's Dialogflow's intent-based framework (Google LLC, Mountain View, CA, USA), that is operationalised to follow the structure of Baba Chan's textbook. The chatbot's conversational flow was built directly around the organisation, learning objectives, and progression of the textbook chapters, allowing it to reflect familiar classroom teaching practices. These intents support key learning activities such as guided lessons, listening and translation exercises, quizzes, badge progression, and navigation across topics. When a learner selects an activity, ChakapBot follows a predefined instructional pathway aligned with the textbook and draws on validated learning materials stored in a curated database. This design helps ensure consistency and reliability, while extending classroom-based instruction into an interactive digital format.

ChakapBot also includes a dictionary lookup function based on Baba William Gwee's dictionary. This function serves as a lexical reference tool: when learners ask about the meaning of a Baba Malay word, the system retrieves the relevant entry from the digitised dictionary. Where available, responses provide English glosses together with linguistic or cultural information drawn from the original dictionary, such as usage context, semantic relationships, or word origins.

As Baba Malay is primarily learned and used through speech, ChakapBot includes a second module in the form of a web-based recording application for creating audio resources. This recording process also supports language documentation by capturing spoken forms of Baba Malay in a validated and reusable format. Through the application, community voice talents record pronunciations, phrases, idioms, and short conversational examples linked to textbook content and dictionary entries. To ensure inclusivity, both male and female voices were recorded. In total, 8714 recordings were produced by each

speaker. All recordings are stored in a cloud repository and undergo a human review process, in which a heritage expert checks pronunciation, intonation, and alignment with written forms. Only approved recordings are made available in ChakapBot. These audio resources are then used across learning activities such as dictionary lookup and listening practice, helping learners connect spoken and written forms of the language.

In addition to these structured, intent-driven interactions, ChakapBot uses generative AI to support more open-ended language queries that cannot be fully anticipated in advance, such as requests for explanations, short narratives, or metalinguistic discussion (Patil & Gudivada, 2024; Qin et al., 2025). Additional resources such as the first author's grammar which was generated through language documented are incorporated into the system to better support these interactions. Generative responses are produced using a Google Gemini 2.5 Flashmodel, but generation is tightly controlled through a retrieval-augmented generation (RAG) pipeline (P. Lewis et al., 2020) deployed on Google Cloud Vertex AI.

The preparation, governance, and ongoing refinement of this RAG-based functionality are managed through ChakapBot's third module, the AI Training Hub. This internal system allows the development team to curate Baba Malay resources, manage metadata, and oversee updates to the retrieval corpus. Through the AI Training Hub, community-approved materials are systematically prepared for retrieval, ensuring that generative responses remain grounded in documented language resources and aligned with community governance principles.

All Baba Malay resources used in ChakapBot are stored in a closed corpus and prepared in advance for retrieval. Source materials are first converted into plain-text formats that allow the system to search and retrieve them efficiently. During this preparation stage, the materials are organised and enriched by identifying key themes and relationships between words and concepts. The corpus is grouped into thematic categories such as family, food, greetings, and home, using the Llama-SEA-LION-v3.5-8B-R model, which was chosen for its stronger linguistic and cultural grounding in Southeast Asian languages. This model is used only at the preparation stage to support culturally appropriate organisation of the materials. The resulting thematic labels, word relationships, and related vocabulary groupings are stored as metadata alongside each entry, helping the system retrieve more relevant and coherent information during use.

All curated texts are segmented using a default fixed chunking strategy, with segments of fault approximately 1024 tokens and an overlap of 256 tokens between adjacent chunks. This configuration was adopted to balance linguistic coherence with computational efficiency. While different types of content in our source materials could in principle be segmented and retrieved separately (Zhao et al., 2025; Nguyen et al., 2025), this pilot implementation retained them within shared chunks, allowing content to be surfaced based on semantic similarity for faster retrieval. All text chunks are indexed using a multilingual embedding model (text-multilingual-embedding-002) and retrieved through similarity-based k-nearest-neighbour matching. Default settings provided by the RAG pipeline were intentionally retained at this stage to support cost effectiveness and stable system performance during the pilot deployment. At the same time, the RAG pipeline remains under active refinement, with ongoing efforts to improve retrieval accuracy, contextual relevance, and response quality as the corpus expands.

Generative responses are produced using a Google Gemini 2.5 Flash model and are constrained through this retrieval-augmented generation (RAG) pipeline, ensuring that outputs are grounded exclusively in curated, community-approved materials. No external data sources or APIs are connected. In addition to retrieval constraints, a system-level prompt governs how the generative model formulates responses. The prompt defines

ChakapBot's role as a supportive, community-oriented interlocutor. It instructs the model to prioritise accuracy, cultural authenticity, and alignment with the curated knowledge base; to model appropriate Baba Malay usage through examples; and to communicate in a warm, learner-friendly manner. Importantly, when a learner's query falls outside the available knowledge base, the prompt requires ChakapBot to state transparently that the information is unavailable, rather than attempting to infer or fabricate content. Together, these retrieval and prompt-based constraints function as ethical and pedagogical guardrails that shape both the content and tone of generative responses.

Prior to deployment, these intent flows and instructional pathways underwent user acceptance testing (UAT) involving the core development team and community collaborators. Altogether 15 testers evaluated the clarity of prompts, appropriateness of instructional sequencing, and accuracy of retrieved content against textbook expectations. Feedback from these cycles informed iterative refinements to intent logic, response phrasing, and activity flow before the pilot rollout. The resultant chatbot resides on an instructional web-based app that is optimized for use on mobile phones.

2.2. ChakapBot's User Interface

The following demonstrates the user interface of ChakapBot (see Figure 1).

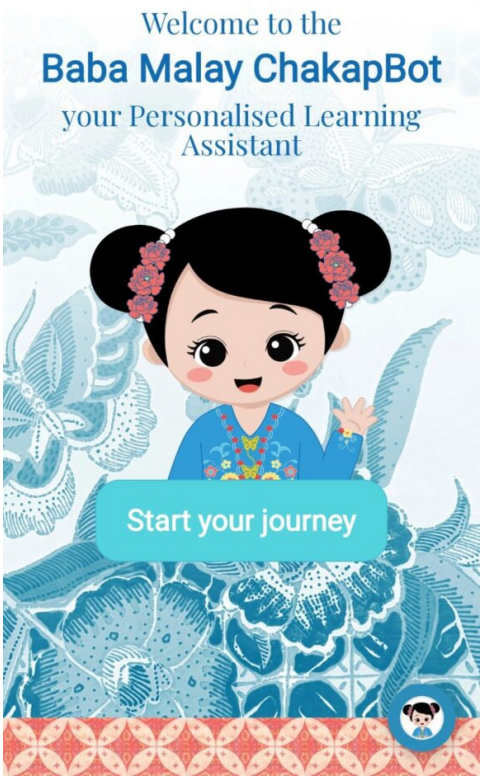

Figure 1. Landing page of ChakapBot.

Notably, ChakapBot is not a general chatbot by way of ChatGPT or Google Gemini. Rather, it is an instructional app that heavily utilises the chat component for learning, with the bot the acting in the role of a "teacher". When users log onto ChakapBot, they are immediately met with "conversation". The bot begins by asking users how they are doing, and users click through to get to the lesson that they want to be at (Figure 2).

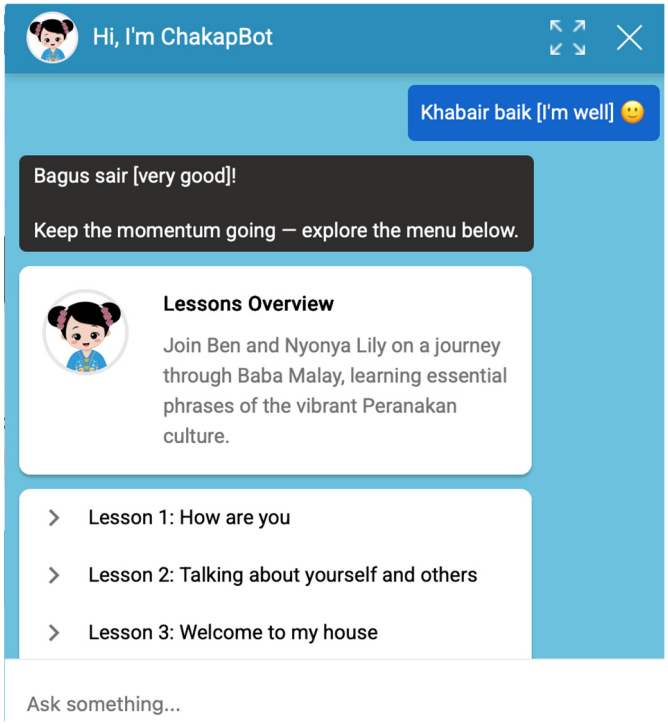

Figure 2. Initial interactions with ChakapBot.

As earlier-mentioned, users engage in scenario-based conversations and complete quizzes to earn badges, while also practising listening, translation, and pronunciation with ChakapBot. The system additionally enables users to record and replay their own speech. Conversations as well as examples for the app were recorded by Baba Malay community members, Baba Lee Yong Ming and Nyonya Carolyn Lim (Figure 3).

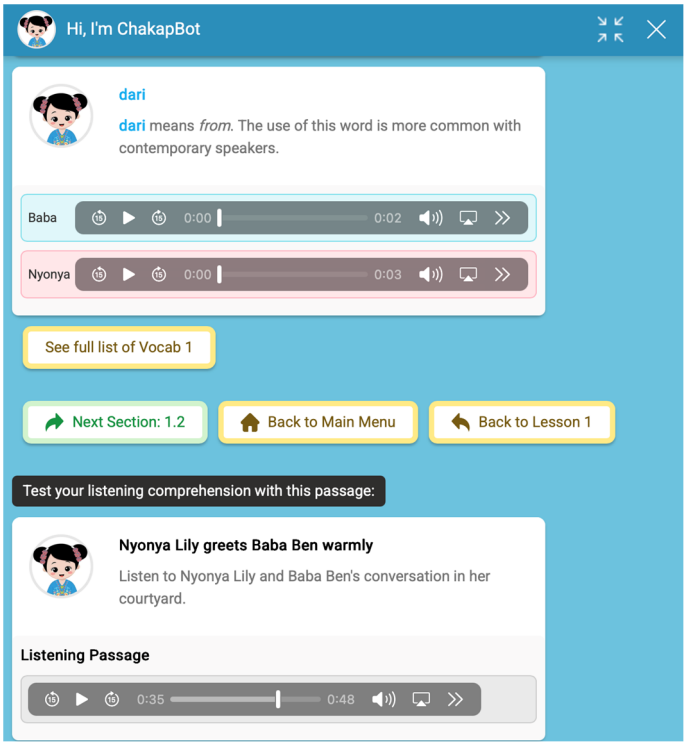

Figure 3. Recorded examples and conversations on ChakapBot.

Alongside structured chatbot-based language learning, users can engage ChakapBot more freely by asking additional questions about the language, drawing on its GenAI functionality. For example, one can type in a query, such as "what is red in Baba Malay?" and receive an appropriate response (see Figure 4).

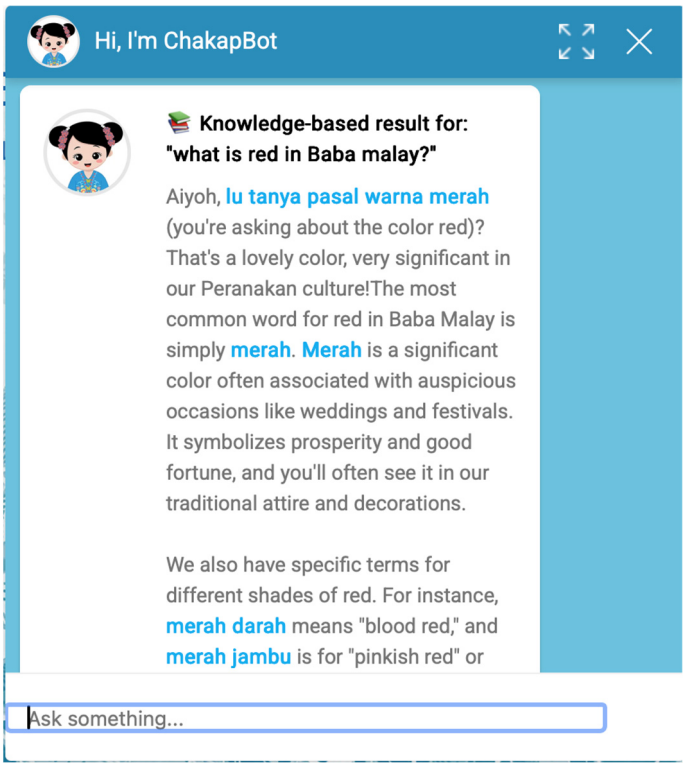

Figure 4. Response to "What is red in Baba Malay?".

Conversely, a user can also ask, what a Baba Malay word means in English, get ChakapBot to recite a story with translation (see Figure 5), or ask a metalinguistic question about the language itself, such as "what is refined in the language?" (See Figure 6).

Relatedly, ChakapBot, being prompted to offer information strictly from its knowledge base, does not profess to exhaustively have all the information there is on Baba Malay. When asked for something that is not within its knowledge base, ChakapBot is trained to tell users that it does not have that information within its knowledge base (see Figure 7).

While the team strives for accuracy as much as possible, in recognition of the limitations of AI-powered tools, ChakapBot includes a disclaimer at the end of each interaction, informing users that the system is AI-powered and advising them to verify important information with trusted sources (see Figure 8).

It is also important to note that ChakapBot is not constructed as a monolingual Baba Malay chatbot. The RAG pipeline constrains ChakapBot's outputs to a curated, community-validated corpus, and this constraint operates primarily at the level of content retrieval. Most of the resources including the textbook that forms its core are bilingual. Generating novel Baba Malay prose carries a meaningful risk of error, and given the project's ethical commitment to accuracy and community ownership, a monolingual chatbot was not deemed appropriate at this stage. From a pedagogical standpoint, the participant profile (predominantly English-dominant young adult learners with limited prior exposure and disrupted intergenerational transmission) made English-mediated instruction a more suitable entry point.

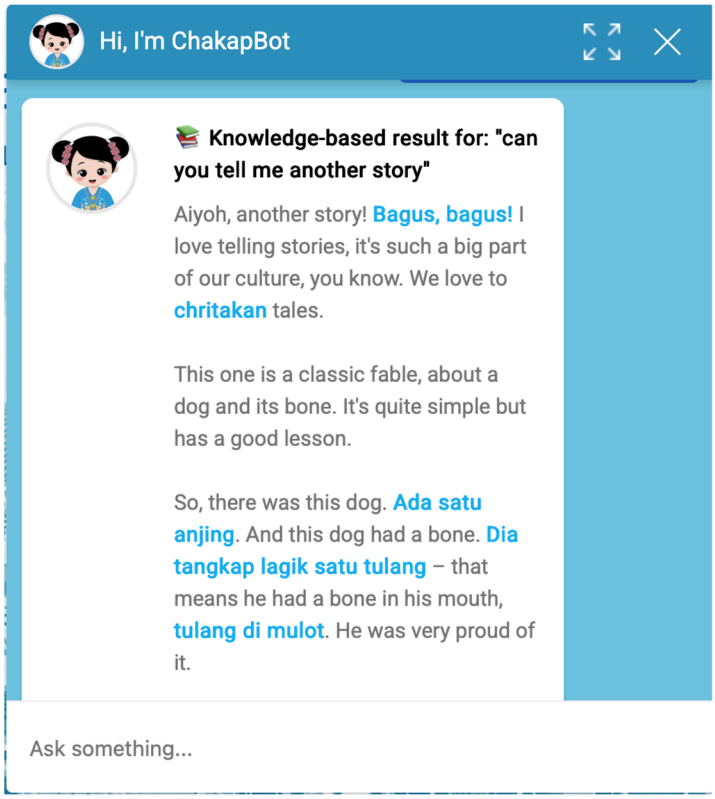

Figure 5. Storytelling with ChakapBot.

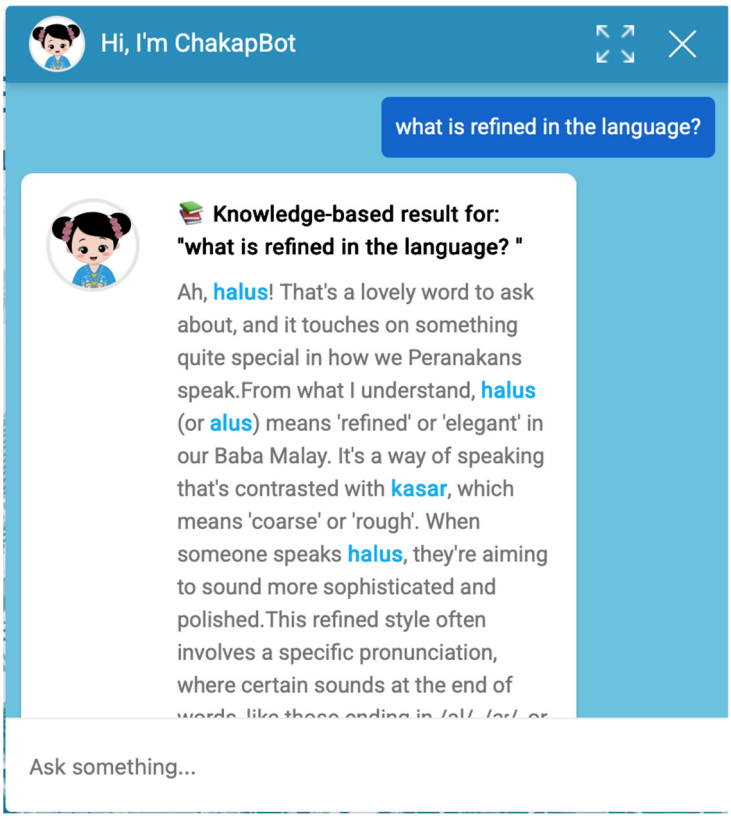

Figure 6. Metalinguistic question on ChakapBot.

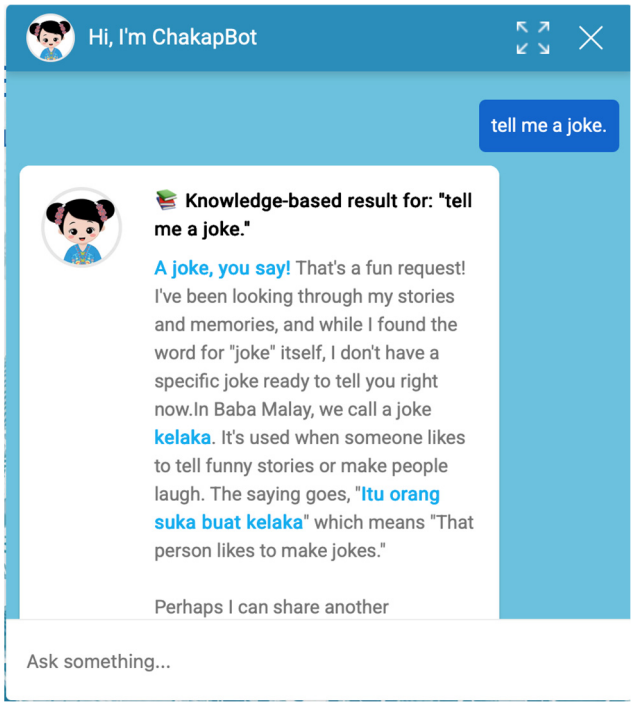

Figure 7. An example of a "null" reply on ChakapBot.

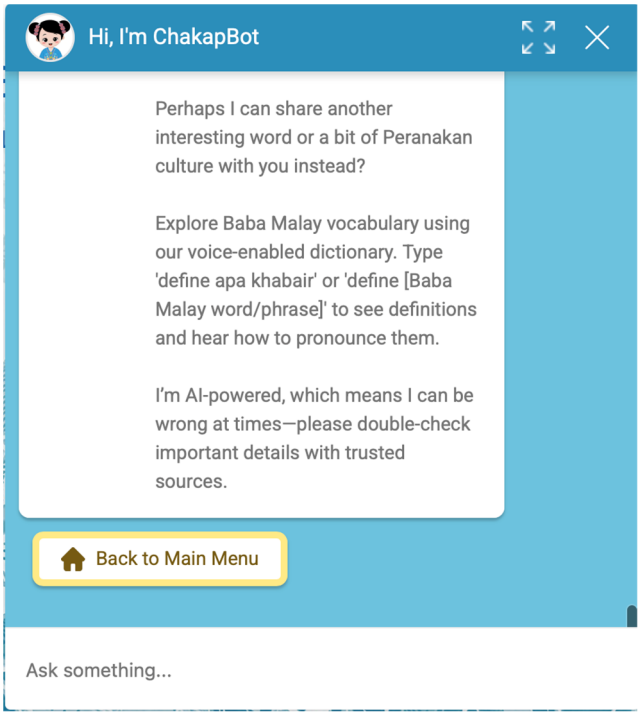

Figure 8. Disclaimer on ChakapBot.

In a pilot project, ChakapBot was used as a tool in a blended learning format together with live classroom lessons. Four lessons were conceived using material and a syllabus that had previously been used in previous iterations of Baba Malay classes (Chan, 2018). The first and third lessons would be delivered in person by Baba Kenneth Chan, a seasoned Baba Malay teacher, while the second and fourth lessons would take place mainly on the ChakapBot app. During the first and third week, learners could also revise the lessons on

the ChakapBot app by themselves. Pre- and post-surveys were also conducted with learners, with questions capturing demographic information and their experiences with ChakapBot.

3. Results

3.1. Pre-Pilot Survey: Participant Demographics, Motivation, and Concerns

The six-week project was piloted with interested public and members of the Peranakan Chinese community between the ages of 10 and 40 years old, most of whom self-reported low levels of proficiency in Baba Malay. Recruitment was carried out primarily through the social media platforms of the Peranakan Association Singapore. Although 40 individuals initially signed up for the programme, 26 participants completed the full pilot and the associated pre- and post-surveys. Participants received remuneration in the form of vouchers upon completion of the study. Notably, only three out of the 26 participants identified themselves as being ethnically Peranakan, although eight reported having been exposed to Baba Malay during childhood. The age and gender distributions of the participants are represented in Figure 9.

Age range	Female	Male	Total
7–12	0	1	1
13–20	3	1	4
21–30	5	7	12
31–40	6	3	9
Total	14	12	26

Figure 9. Age and gender profile of participants.

The age distribution indicates that ChakapBot primarily engaged younger adults, plausibly suggesting the potential effectiveness of digital chatbot-based tools for adult learners. There was poorer engagement among younger participants, which may be attributable to the highly structured schedules of young children in Singapore, who often balance school, tuition, co-curricular, and enrichment activities (Lim, 2024), leaving relatively little discretionary time for participation in initiatives such as the one described here.

Participants signed up for the programme for a range of reasons, including a desire to learn or reconnect with Baba Malay and the Peranakan Chinese culture. Qualitative responses reflected perspectives from both members of the public and participants who identified with the Peranakan Chinese heritage, as illustrated in the excerpts below:

> "I was interested in trying a new language."
>
> "Pick it up for fun and understand the culture."
>
> "…it's important for me to reconnect with my heritage and the language of Baba Malay, my ancestors, is integral to that."
>
> "Finding the heritage in my family. Understand that my great-grandmother is a peranakan and wanted to be closer to my roots."

Responses also indicated that, for some participants, engagement was driven by explicit concerns about language endangerment:

> "…I am interested in investigating the conversation surrounding preservation vs. conservation, why is it worth it to preserve a culture and how we might preserve traditions."

> "Think it would be a good way to bond with the maternal side of my family and to preserve the language that I never got to learn."

Notably, while none of the participants identified the AI chatbot component itself as their primary reason for signing up for the programme, the data indicate a clear openness to digital learning. In response to a question on preferred instructional formats or learning environments, only four of the 26 participants expressed a preference for traditional in-person classroom instruction, textbooks, or printed materials. The remaining participants indicated preferences for digital modalities, including online classes, webinars, digital learning platforms, mobile applications, and virtual (AI-supported) tutoring. Furthermore, pre-pilot survey responses showed that 17 participants were willing to dedicate at least 30 min per week to using ChakapBot for self-directed learning, while the remaining participants indicated a willingness to spend between 15 and 30 min per week. This suggests that the flexible, self-paced nature of ChakapBot may be particularly well suited to learners with busy schedules, supporting sustained engagement with Baba Malay.

Additionally, although only seven participants expressed concerns about the use of AI, their comments are nonetheless noteworthy. These concerns centred primarily on the accuracy of ChakapBot, given its AI-driven nature:

> "Authenticity."

> "How to ensure that what I am learning is correct."

Ethical concerns that were raised include:

> "mainly the invasion of privacy (afterall it is giant data collection)."

> "will this form of reinforcement learning help the AI to proceed towards AGI and ASI?"

Overall, the values articulated by participants—particularly transparency and ethical safeguards—align closely with ChakapBot's constrained, community-curated design.

3.2. Post-Pilot Survey: Where ChakapBot Succeeds and Areas for Improvement

While the conclusions that can be drawn from a six-week pilot study are necessarily limited, the findings nonetheless allow for useful observations regarding ChakapBot within the context of blended learning.

In general, ChakapBot was positively received by participants. Most participants reported that they were able to spend at least 15 min on the app weekly (see Figure 10).

Only seven participants reported spending less than 15 min on ChakapBot, with fewer engaged for more than an hour. Many of the participants engaged with the tool in manageable sessions of between 15 and 30 min.

Most participants reported that they used ChakapBot with relative ease (see Figure 11) with none of the participants rating the tool 0 for "Not easy at all".

Additionally, participants reported that they have seen improvement in their Baba Malay, with none of the participants selecting 0 (see Figure 12). About half the participants reported substantial improvement, rating their perceived improvement as 4 or 5 on the Likert scale.

When asked how much they would attribute their progress to ChakapBot, distribution centred on mid-high points of the Likert scale, with larger number of participants stating that they benefited "Somewhat" and "Quite a lot" from use of the app (see Figure 13).

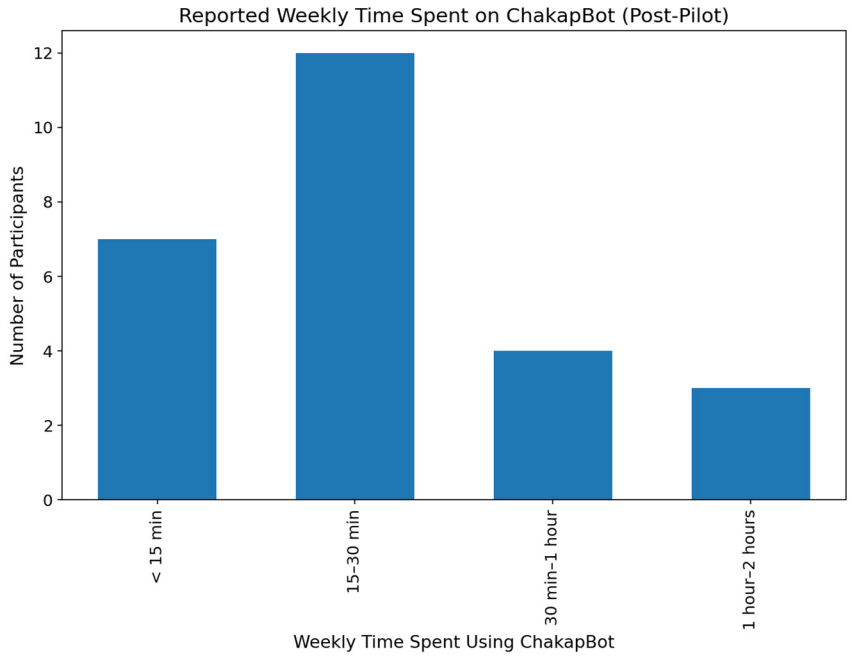

Figure 10. Reported weekly time spent on ChakapBot.

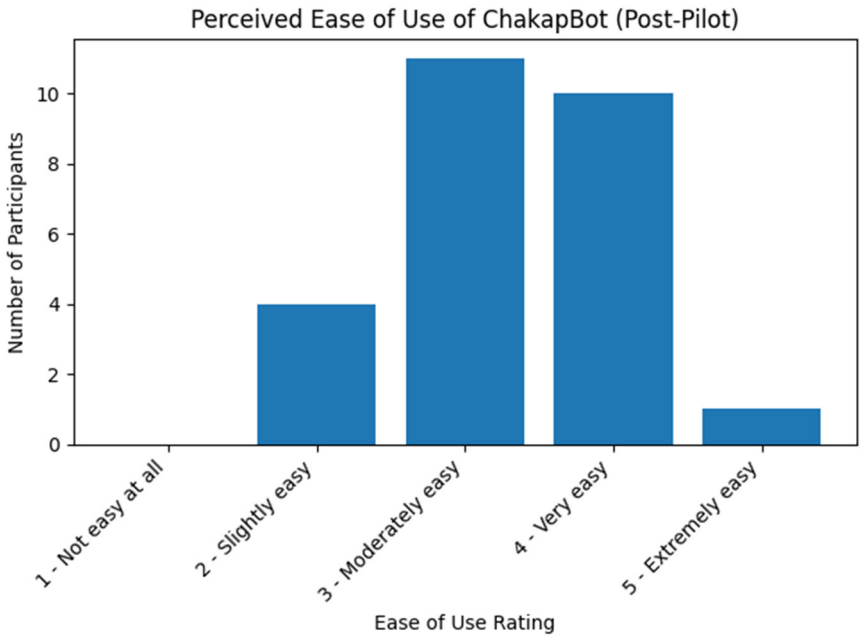

Figure 11. Perceived ease of use of ChakapBot.

The participants' self-reported attribution of progress to ChakapBot can be compared with how they attributed progress to in-person classes. Participants were more inclined to attribute "Quite a lot" or "A lot" of their progress to the in-person classes. In this case, the distribution is clearly more strongly weighted towards the higher end of the Likert scale (see Figure 14).

ChakapBot was perceived as an important supporting component in the blended learning environment, contributing meaningfully, but it was not perceived as the sole or primary driver of progress in language learning. Rather, the in-person classes featuring a respected Baba Malay community member as teacher was perceived to have been a crucial part of the learning experience. Learning was experienced by these students as being

blended and complementary, rather than being driven by any single modality, given that the rating for ChakapBot was not skewed to the lower end of the scale either.

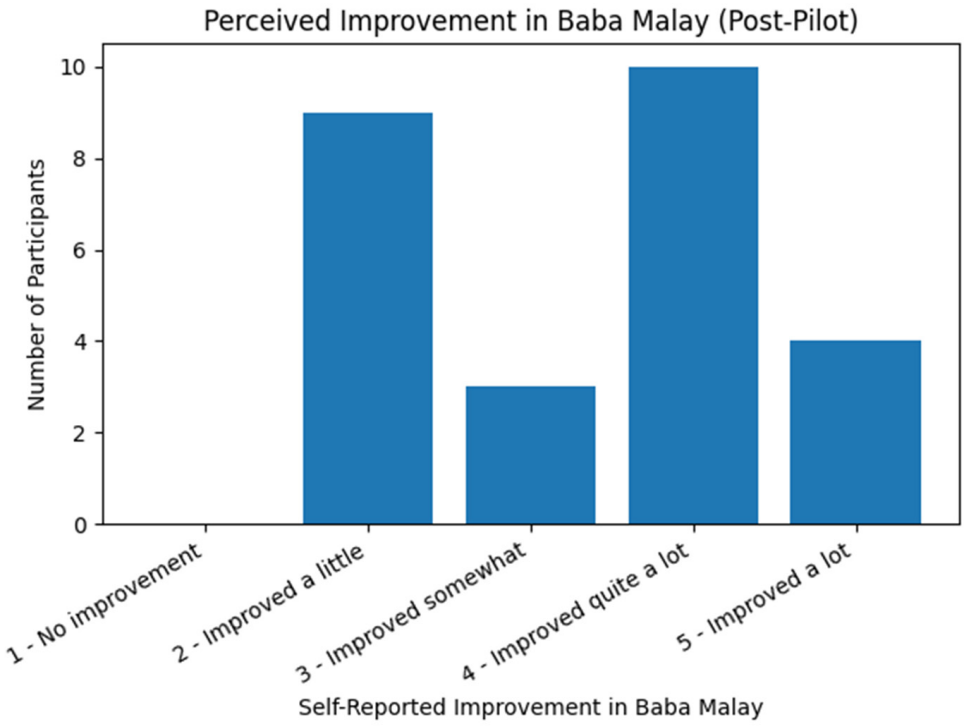

Figure 12. Self-reported improvement in Baba Malay.

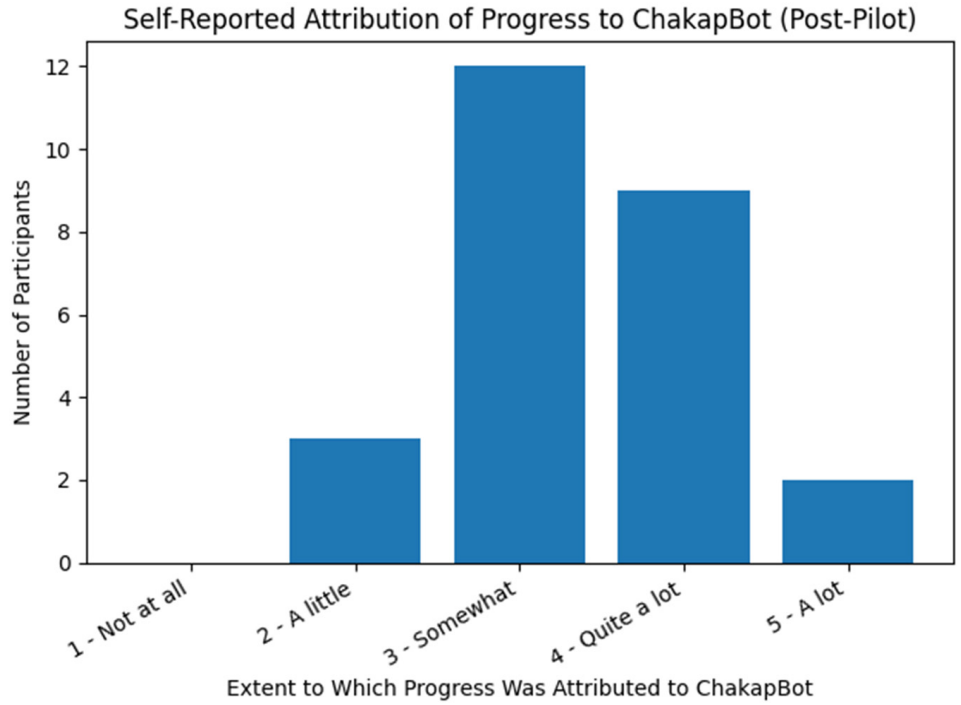

Figure 13. Self-reported attribution of progress to ChakapBot.

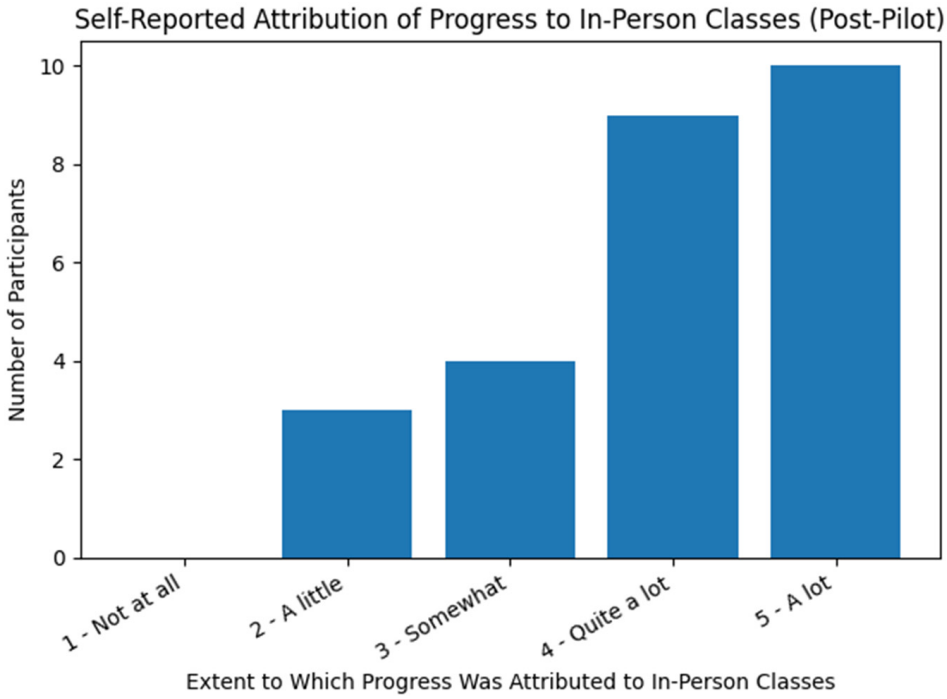

Figure 14. Self-reported attribution of progress to in-person classes.

Crucially, on an open-ended question about what participants liked about ChakapBot, participants' responses clustered on a few recuring themes. Responses to the question "What did you like about ChakapBot?" indicate that participants particularly valued its ease of use, flexibility, and low-pressure learning environment. Many highlighted the convenience of practising outside of class, the opportunity for reinforcement and discovery, and the engaging, interactive nature of the chatbot format, positioning ChakapBot as a useful complementary learning support:

"The app was easy and intuitive to use."

"I can learn anytime anywhere."

"I like that it is available 24/7, and is a good source of vocabulary or sentences that may not be found in the textbooks or covered during classes."

"It has audio function which has been really helpful for learning pronunciation."

"The AI feature that allows us to pose any question and it will answer to the best of its availability just like the likes of Chatgpt and Copilot. . . .I also appreciate the voice recordings as it accurately captures how the language should be spoken from an expert POV."

"the LLM structure where I can spontaneously send question prompts to clarify my queries/doubts."

Additionally, participants were also asked what they did not like about ChakapBot. Responses to this question primarily highlighted the lack of a speech recognition system in the specific case of ChakapBot, other limitations typical of AI-mediated tools, such as constraints in conversational depth, occasional uncertainly about accuracy, and constraints in conversational depth, highlighting in part the desire for expanded content.

"It had no speak-to-check function or function for us to ask questions."

"how the efficiency decreases when further prompting is given."

> It's still not comprehensively trained to answer every kind of questions related to Baba Malay, like list me all words that have alus form."
>
> "ChakapBot could have more interactive element when prompted, e.g., games, videos to add to the learning and understanding culture. Focusing not just on language."

Participants were also asked, "what did you feel you could learn better in the in-person classes?" Their responses highlight the perceived strengths of human-led instruction, particularly in conveying linguistic nuance, cultural context, and pragmatic usage. Participants also emphasised the value of immediate clarification, tailored explanations, feedback, and opportunities for interaction with instructors and peers:

> "To understand social nuances and culture."
>
> "The speaking portion. Hearing the organic use of the language made picking up the intonation and nuances of the language easier."
>
> "I could practice speaking Baba Malay and get an active feedback. . . "
>
> "More Engagement with the actual speakers of the language."

Considered together, responses underscored the complementary roles of the instructional AI chatbot and the traditional classroom for learning Baba Malay, with the traditional classroom serving as the main entry point into the language, and the instructional chatbot sustaining the learners' interest and further engaging them outside the classroom.

Overall, the ChakapBot pilot demonstrates promising outcomes in a language revitalisation context. Even within the limited six-week duration, participants reported finding opportunities to use Baba Malay beyond the classroom. Specifically, 16 of the 26 participants indicated that they had spoken Baba Malay outside of class, including with family members, Peranakan friends, colleagues in the workplace, and at Peranakan community events.

The ChakapBot pilot programme being six weeks long is too short for any meaningful consideration of proficiency gains as an indicator of revitalisation success—instead its real success can be articulated in terms of encouraging the use of Baba Malay in the everyday lives of the participants.

4. Discussion: Implications of an AI-Powered Chatbot for Language Revitalisation and Language Documentation

The ChakapBot pilot highlights the potential of ethically constrained, community-led AI tools for language revitalisation, particularly among young adult learners. Digital tools such as AI-powered chatbots may be especially appealing to learners with limited prior exposure to the language—often adults for whom intergenerational transmission has been disrupted and who are encountering the language later in life. For such learners, affective barriers, including anxiety, fear of error, and reduced self-confidence, have been shown to impede second-language learning (Cohen & Norst, 1989; Schumann, 1975). Engagement with an AI-powered chatbot may help mitigate these affective barriers by providing a low-stakes, non-judgemental environment for practice, even if learners do not explicitly articulate this effect in qualitative feedback. In addition, chatbot-based tools are well suited to adult learners managing work and family commitments, as they offer flexible, on-demand opportunities for language use across time and settings.

The findings further suggest that chatbot-based applications are most effective when embedded within a blended learning model that combines human-led instruction with flexible opportunities for independent practice. Results from the pilot are encouraging, with effectiveness reflected in participants' reported use of Baba Malay beyond the classroom, including in domains where they would not traditionally speak the language, such as at

home, with friends, and at Peranakan community events. While human-led instruction—particularly when delivered by a respected community teacher—supports linguistic nuance, corrective feedback, and appropriate cultural grounding, the AI chatbot facilitated practice, reinforcement, and sustained engagement outside formal instructional settings. These findings indicate that blended learning models are more effective than standalone technological solutions in this context. Importantly, AI is not intended to usurp the role of community experts, especially in cases where the language is endangered but not dormant. Rather, as Ingram (2025, p. 73) notes, AI can be used effectively "as a complement to, rather than a replacement for, traditional methods of cultural and linguistic preservation."

Additionally, the ChakapBot pilot demonstrates that GenAI can be meaningfully used in language revitalisation when it is carefully and effectively constrained. The benefits of GenAI, including a smoother and more responsive user experience, can be realised without sacrificing linguistic accuracy or cultural integrity when generation is grounded in community-curated documentation and governed by clear ethical and pedagogical boundaries. In this way, GenAI functions not as an autonomous knowledge producer, but as a facilitative layer that helps learners access documented language resources in more flexible and engaging ways, supporting practice and sustained engagement while remaining accountable to community authority. Although this implementation deliberately constrains the chatbot's responses to a curated knowledge base—an essential design choice given the emphasis on ethics, authenticity, and accuracy in revitalisation and documentation work—the pilot represents an initial step rather than a final system. The findings point to clear directions for future development, including the expansion of the curated corpus, the integration of speech-based functionalities, and longer-term deployment to examine sustained engagement and learning outcomes over time.

Beyond its use as a pedagogical tool, ChakapBot shows how AI systems can operate at the interface of language documentation and revitalisation. On the one hand, the chatbot actively draws on existing documentation—such as the dictionary, textbook, and grammatical description of Baba Malay—thereby operationalising documented knowledge and making it accessible for everyday learning and use. On the other hand, ChakapBot contributes to language documentation efforts through its audio recording and validation workflow, which captures spoken forms of the language in a structured, community-governed format. By integrating new, validated audio data alongside previously documented resources, the system supports a process in which documentation informs revitalisation and revitalisation generates new forms of documentation.

The ChakapBot pilot illustrates how AI-supported tools can contribute meaningfully to both language revitalisation and language documentation when they are developed in close collaboration with the communities they are intended to serve. The findings underscore the importance of community participation and curation in safeguarding linguistic and cultural integrity, as well as the importance of addressing ethical concerns surrounding AI use. Participants engaged with ChakapBot are reflective and critical users, and transparency and clear boundaries around such system capabilities are key to fostering the trust and sustained engagement of community members. More broadly, this study shows that when generative AI is carefully constrained and grounded in community-led documentation, it can function not as a replacement for human expertise, but as an infrastructure that connects documentation, teaching, and revitalisation in sustainable ways.

While the study was conducted to capture more instrumental outcomes, future studies like this one can also capture the affective and aspirational dimensions of language revitalization. Looking ahead, another important question for future research concerns the degree to which AI tools can meaningfully shift language vitality indicators, and not merely learner engagement metrics. A longer-term study would be more useful in this

regard. Additionally, there is also the question of whether the technology can at some point in the future function less as a teacher, but more as a social infrastructure—one that recreates a sense of community that is gradually being lost. Such questions are beyond the scope of this study but are interesting ones that should be considered in the future.

Author Contributions: Conceptualization, H.Z. and N.H.L.; methodology, H.Z. and N.H.L.; validation, H.Z.; formal analysis, H.Z. and N.H.L.; investigation, H.Z. and N.H.L.; resources, H.Z.; data curation, H.Z. and N.H.L.; writing—original draft preparation, N.H.L. and H.Z.; writing—review and editing, N.H.L. and H.Z.; visualization, N.H.L.; project administration, H.Z.; funding acquisition, H.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Heritage Board, Singapore through the Heritage Research Grant, grant number HRG-60.

Institutional Review Board Statement: This study was approved by the Temasek Polytechnic Ethics Committee on 21 October 2024 (study code: SR35-IIT-2024_ER) and conducted in accordance with institutional requirements. All data were collected and managed in accordance with ethical guidelines, with no personally identifiable information included in the analysis or reporting.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy and ethical restrictions.

Acknowledgments: The authors are also grateful to the Peranakan Association Singapore for its support and encouragement of this initiative, and for helping to connect the project team with members of the Peranakan community.

Conflicts of Interest: The authors declare no conflict of interest.

References

Ansaldo, U., Lim, L., & Mufwene, S. S. (2007). The sociolinguistic history of the Peranakans: What it tells us about "creolization". In U. Ansaldo, S. Matthews, & L. Lim (Eds.), *Deconstructing creole. Typological studies in language. 73* (pp. 203–226). John Benjamins.

Barath, H. (2025, April 9). *Language preservation efforts gets an AI boost*. Darthmouth University. Available online: https://home.dartmouth.edu/news/2025/04/language-preservations-efforts-get-ai-boost?utm_source=chatgpt.com (accessed on 1 February 2026).

Basar, E., Balaji, D., He, L., Hendrickx, I., Krahmer, E., De Bruijn, G., & Bosse, T. (2023, July 19–21). *HyLECA: A framework for developing hybrid long-term engaging controlled conversational agents*. Proceedings of the 5th International Conference on Conversational User Interfaces, Eindhoven, The Netherlands.

Blanks, K. (n.d.). *Linguistic revival: How Japan restored the native Ainu language with "AI Pirika"*. Rewired. Available online: https://stanfordrewired.com/post/japan-restored-ainu-ai-pirika (accessed on 1 February 2026).

Brewer, G. L. (2022, June 3). Lakota elders helped a white man preserve their language. Then he tried to sell it back to them. *NBC News*. Available online: https://www.nbcnews.com/news/us-news/native-american-language-preservation-rcna31396 (accessed on 1 February 2026).

Brixey, J., & Traum, D. (2020). Masheli: A Choctaw-English bilingual chatbot. In *Proceedings of the 11th international workshop on spoken dialog systems*. IWSDS.

Chan, K. (2018). *Mari Chakap Baba: A comprehensive guide to the Baba Nyonya language*. Gunong Sayang Association.

Chia, F. (1983, August 19). *Peeping through the nonya window curtain and seeing the decline of the Baba culture*. A Talk Held at the Central Library, National University of Singapore.

Cohen, Y., & Norst, M. J. (1989). Fear, dependence and loss of self-esteem: Affective barriers in second language learning among adults. *RELC Journal*, *20*(2), 61–77. [CrossRef]

Dasgupta, D. (2025, October 7). India launches AI app to save its many tribal languages before they vanish. *The Straits Times*. Available online: https://www.straitstimes.com/asia/south-asia/india-launches-ai-app-to-save-its-many-tribal-languages-before-they-vanish (accessed on 1 February 2026).

Department of Statistics Singapore. (2020). *Singapore census of population 2020, statistical release 1: Demographic characteristics, education, language and religion*. Available online: https://www.singstat.gov.sg/publication-resources/singapore-census-of-population-2020-statistical-release-1-demographic-characteristics-education-language-and-religion (accessed on 5 February 2026).

Glorieux-Stryckman, M. (2024, December 16). AI outrage: Error-riddled Indigenous language guides do real harm, advocates say. *The Gazette*. Available online: https://montrealgazette.com/news/ai-outrage-error-riddled-indigenous-language-guides-do-real-harm-advocates-say (accessed on 1 February 2026).

Gwee, W. T. H. (1993). *Mas sepuloh: Baba conversational gems*. Armour Publishing.

Gwee, W. T. H. (2006). *A Baba Malay dictionary: The first comprehensive compendium of straits Chinese terms and expressions*. Tuttle Publishing.

Ingram, M. (2025). The role of AI in language preservation and revitalization. In *AI for community* (pp. 55–80). Chapman and Hall/CRC.

Jain, R. (Ed.). (2021). *Multilingual Singapore: Language policies and linguistic realities*. Routledge.

Jampel, S. (2025, July 31). Can A.I. help revitalize indigenous languages? *Smithsonian Magazine*. Available online: https://www.smithsonianmag.com/science-nature/can-ai-help-revitalize-indigenous-languages-180987060/ (accessed on 1 February 2026).

Lau, A. G. L. (1984). *Language death with particular reference to the Baba Malay of Singapore* [Master's thesis, University of York].

Lau, J. (2025, June 26). *Can AI speak the language Japan tried to kill?* BBC. Available online: https://www.bbc.com/future/article/20250625-can-ai-speak-the-language-japan-tried-to-kill (accessed on 1 February 2026).

Lee, N. H. (2014). *A grammar of Baba Malay with sociophonetic considerations* [Ph.D. dissertation, University of Hawai'i].

Lee, N. H. (2019). Peranakans in Singapore: Responses to language endangerment and documentation. *Language Documentation & Conservation*, *19*, 123–140.

Lee, N. H. (2022). *A grammar of modern Baba Malay*. Mouton de Gruyter.

Lee, N. H., & Van Way, J. (2018). Assessing degrees of endangerment. In K. Rehg, & L. Campbell (Eds.), *Oxford handbook of endangered languages* (pp. 48–65). Oxford University Press.

Lewis, M. P., & Simons, G. F. (2010). Assessing endangerment: Expanding Fishman's GIDS. *Revue Roumaine de Linguistique*, *55*(2), 103–120.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in neural information processing systems* (Vol. 33, pp. 9459–9474). Curran Associates, Inc.

Lim, C. (2024, June 16). Commentary: Are Singapore children overscheduled during school holidays? *Channel News Asia*. Available online: https://www.channelnewsasia.com/commentary/kid-school-june-december-holiday-activities-play-benefits-parenting-4410586?utm_source=chatgpt.com (accessed on 5 February 2026).

Long, Y., & Zhang, Y. (2024). Enhanced classroom dialogue sequences analysis with a hybrid AI agent: Merging expert rule-base with large language models. *arXiv*. [CrossRef]

Mila—Quebec Artificial Intelligence Institute. (n.d.). *First languages AI reality*. Available online: https://mila.quebec/en/ai4humanity/applied-projects/first-languages-ai-reality (accessed on 1 February 2026).

Nguyen, H., Nguyen, T., & Nguyen, V. (2025). Enhancing retrieval-augmented generation with hierarchical text segmentation chunking. *arXiv*. [CrossRef]

O'Grady, W., & Hattori, R. (2016). Language acquisition and language revitalization. *Language Documentation & Conservation*, *10*, 45–57.

OpenAI Press Release. (2025, March 14). How Iceland is using GPT-4 to preserve its language. *Arctic Today*. Available online: https://www.arctictoday.com/how-iceland-is-using-gpt-4-to-preserve-its-language/ (accessed on 1 February 2026).

Pakir, A. (1986). *A linguistic investigation of Baba Malay* [Ph.D. dissertation, University of Hawai'i].

Patil, R., & Gudivada, V. (2024). A review of current trends, techniques, and challenges in Large Language Models (LLMs). *Applied Sciences*, *14*(5), 2074. [CrossRef]

Qin, Y., Zhou, Y., Huang, X., & Huricha, B. (2025). Empowering foreign language education with large language models: Applications and insights. *Journal of Higher Education Teaching*, *2*, 100–101. [CrossRef]

Rappa, A. L., & Wee, L. (2006). *Language policy and modernity in Southeast Asia: Malaysia, the Philippines, Singapore, and Thailand*. Springer.

Schumann, J. H. (1975). Affective factors and the problem of age in second language acquisition. *Language Learning*, *25*(2), 209–235. [CrossRef]

Tan, C. B. (1979). *Baba and Nyonya: A study of the ethnic identity of the Chinese Peranakan in Malacca*. Cornell University.

Te Hiku Media. (n.d.). *Te Hiku Tech*. Available online: https://tehiku.nz/te-hiku-tech/ (accessed on 1 February 2026).

The Steam Connection. (2024). *SkoBots*. Available online: https://www.steamconnection.org/skobots (accessed on 1 February 2026).

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *59*, 433–460. [CrossRef]

UNESCO. (2022, January 5). *Indigenous languages: Gateways to the world*. UNESCO. Available online: https://www.unesco.org/en/articles/cutting-edge-indigenous-languages-gateways-worlds-cultural-diversity?utm_source=chatgpt.com (accessed on 1 February 2026).

UNESCO Ad Hoc Expert Group on Endangered Languages. (2003, March 10). *Language vitality and endangerment. Document adopted by the international expert meeting on UNESCO programme safeguarding of endangered languages*. UNESCO.

Wang, L. (2024). Artificial intelligence's role in the realm of endangered languages: Documentation and teaching. *Applied and Computational Engineering*, *48*, 123–129. [CrossRef]

Wiltshire, B., Bird, S., & Hardwick, R. (2024). Understanding how language revitalisation works: A realist synthesis. *Journal of Multilingual and Multicultural Development*, *45*(9), 3946–3962. [CrossRef]

Zhao, J., Ji, Z., Fan, Z., Wang, H., Niu, S., Tang, B., Xiong, F., & Li, Z. (2025). MoC: Mixtures of text chunking learners for retrieval-augmented generation system. In *Proceedings of the 63rd annual meeting of the association for computational linguistics: Vol. 1, Long Papers* (pp. 5172–5189). Association for Computational Linguistics.