Woollaston et al. *Research and Practice in Technology Enhanced Learning* (2027) 22:3
https://doi.org/10.58459/rptel.2027.22003

Research and Practice in
Technology Enhanced Learning

RESEARCH **Free and Open Access**

TAMMY: Usability and effectiveness of a translation practice chatbot for Japanese EFL learners

Steve Woollaston [1] *, Brendan Flanagan [2], Patrick Ocheja [3], Yiling Dai [4], and Hiroaki Ogata [5]

*Correspondence:
s.m.woollaston@gmail.com
Graduate School of Informatics,
Kyoto University, Kyoto, Japan
Full list of author information is
available at the end of the article

Abstract

This study evaluates the usability and effectiveness of an LLM-powered chatbot, TAMMY (Translation Assistant for MMasterY), designed to support EFL translation practice for Japanese students. Tammy was designed to provide Japanese to English translation tasks, offering feedback and guidance to learners. Using chatlog data and a questionnaire, we assessed Tammy's response validity, its success in guiding learners to accurate translations, changes in English proficiency, and student perceptions of ease of use, usefulness, and friendliness. Findings revealed high chatbot response validity (97% valid) and moderate success in guiding learners to accurate translations (64%). Students perceived Tammy as friendly and easy to use, yet their intention for continued use was neutral. While no significant relationship between Tammy usage and changes in English proficiency was found, our analysis provides insights into the design of AI-integrated translation tools. The results underscore that AI-powered chatbots can serve as effective, non-judgmental assistants for unlimited language practice, but they are most impactful as a complementary tool within a broader pedagogical framework, not a standalone solution. This work offers concrete recommendations for future chatbot design improvements to better align AI capabilities with specific learning objectives, thereby enhancing pedagogical effectiveness and promoting sustained student engagement.

Keywords: chatbot, conversational agents, EFL, translation, usability, chatlogs, exTAM, language learning, GenAI, LLM, CALL

Introduction

The integration of technology into English as a Foreign Language (EFL) education has expanded significantly (Zhang & Zou, 2022), with Computer Assisted Language Learning (CALL) systems playing a pivotal role in enhancing learning outcomes through tools like automated spaced repetition systems (e.g., Anki flashcards) and gamified platforms such

© The Author(s). 2026 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

as Duolingo. Learners and teachers increasingly use CALL in both formal and informal contexts (Levy & Stockwell, 2013). Numerous free and paid apps and tools are available for learning, practising language skills, and connecting with language partners (Bowker et al., 2008). Learning the English language is particularly popular due to its role as the *lingua franca* in international business and academia (Smith, 2005). Chatbots have emerged as a promising tool for language practice and instruction, providing an interactive and engaging dialogue-based platform for learners to gain exposure to natural conversations and improve their skills in the target language. They offer learners opportunities to practise conversational skills and receive real time feedback. While existing research highlights chatbots' effectiveness in grammar and vocabulary acquisition (Jeon, 2023; N.Y. Kim, 2019), their application to translation tasks, an important yet underexplored component of second language acquisition, remains limited. Translation practice fosters language acquisition by strengthening vocabulary retention, highlighting grammar similarities and differences, and promoting metalinguistic awareness (Anderson, 2018; Dagilienė, 2012). However, the specific contributions of chatbot-mediated translation practice to English proficiency have not been extensively explored. This gap underscores the need to investigate chatbots not merely as conversational tools, but as facilitators of structured language practice that may improve language proficiency. Given the rise of LLMs, there is now a unique opportunity to design and evaluate a chatbot that provides nuanced, iterative feedback on complex tasks like translation—a process that differs significantly from the rule-based feedback found in traditional language-learning apps like Duolingo and Memrise.

In this study, we introduce and evaluate TAMMY (Translation Assistant for MMasterY), an EFL translation practice chatbot powered by a large language model (LLM). The chatbot, referred to as Tammy from here on, is designed to support Japanese EFL learners through scaffolded translation tasks. Unlike many general-purpose LLM chatbots or rule-based educational bots, Tammy is specifically engineered to provide iterative, hint-based feedback for translation practice, avoiding immediate full solutions to encourage deeper engagement and processing. Tammy is evaluated using a questionnaire and analysis of chatlog data and English test results.

Related Work

Translation tasks for learning EFL

While translation tasks have experienced shifts in popularity, research consistently highlights their effectiveness in second language learning. From traditional Grammar-Translation Method drills to modern cognitive and communicative approaches, translation remains a core component of language instruction (Antón Remírez & Lázaro Ibarrola, 2011). Translation tasks are effective language learning activities for:

1. **Vocabulary acquisition** (Folse, 2004): Learners are able to connect their L1 (first language) vocabulary knowledge to the L2 (target language) and notice similar words and "false friends" (words that look or sound the same, but have entirely different meanings).
2. **Grammar and syntax**: Translation tasks provide numerous formulaic phrases and help learners understand and apply grammatical structures more accurately (Murtisari et al., 2020).
3. **Reading comprehension**: Translation exercises enhance reading comprehension by promoting close attention to details. (Chang & Hsu, 2011).
4. **Writing proficiency**: Translation tasks improve writing by enhancing accuracy, cohesion, and coherence (Tavakoli et al., 2014) and reducing cognitive load by allowing idea development in L1 before translation (Huh, 2001; Saleh et al., 2021).

Calis and Dikilitas (2012) investigated the use of translation tasks as a learning activity in EFL classes. Through a classroom-based study with elementary EFL learners, the study found that learners generally had positive perceptions towards using translation exercises for learning English. The learners believed that translation helped them improve reading comprehension, memorise vocabulary, and develop writing skills. The study also revealed that learners frequently used strategies like consulting dictionaries and checking their understanding against available translations. The findings suggest that incorporating translation tasks into EFL instruction can promote the development of both receptive and productive language skills by allowing learners to draw connections between the target language and their native language.

Translation tasks offer a multifaceted approach to improving language proficiency by engaging learners across several key areas, or strands as Nation (1996) calls them in his seminal work on effective language learning. They directly support *meaning-focused input* as learners must fully comprehend the source text to accurately translate its ideas and messages into the target language. This process inherently fosters *language-focused* instruction, requiring deliberate attention to specific linguistic features such as grammatical structures, syntax, spelling and punctuation to ensure accurate application and understanding. Furthermore, depending on the direction of the translation, producing the translated text constitutes *meaning-focused output*, pushing learners to actively retrieve and apply their knowledge to convey ideas effectively, thereby enhancing productive language skills.

Translation using CALL

Many language learning apps and digital tools integrate translation tasks into their teaching methods, with recent studies exploring their effectiveness and user experiences. While popular platforms like Duolingo and Memrise use translation in a gamified context to teach vocabulary and grammar, their rule-based, repetitive nature often lacks the nuanced, contextual feedback required for complex tasks. For example, a case study by Loewen et al. (2019) on Duolingo found that while learners improved, their gains were limited, and the app's focus on gamification sometimes overshadowed genuine language acquisition. Similarly, an autoethnographic study by Karjo and Andreani (2018) noted that Memrise's spaced repetition, while memory-based, could be perceived as inefficient and boring. These tools are often criticised for teaching language skills incompletely without deeper explanations. In contrast, Babbel and Google Translate offer a slightly different approach, with Babbel providing more explicit grammar instruction and Google Translate serving as a tool for vocabulary and paraphrasing. A study by Kessler, Loewen, & Gönülal (2023) showed that Babbel users perceived it as more effective for grammar and speaking than Duolingo, yet both apps still faced issues with user motivation and persistence. A qualitative study by Inderawati et al. (2023) on Google Translate noted its benefits for vocabulary and grammar, but also pointed out its limitations in producing natural sentence structures and the risk of over-reliance.

The advent of Generative AI (GenAI) has fundamentally changed the landscape of language learning by offering a more flexible and adaptable approach than these traditional platforms. Unlike the rule-based feedback of Duolingo or the memory-based repetition of Memrise, LLMs can provide nuanced, iterative feedback and handle errors in a way that mimics human-like interaction. Recent studies have begun to explore GenAI's potential in language learning. For instance, Xu, Sun, and Hu (2025) found that while GenAI effectively addresses linguistic conventions in translation, it still requires instructor oversight for cultural nuances. Similarly, Duolingo's recent integration of GenAI has shown positive impacts on learner engagement and communication skills (Ouyang et al. 2024; Phanwiriyarat et al. 2025). However, these studies also noted limitations for intermediate and advanced language learners.

Chatbots for Language Learning

Chatbots gained prominence in user support for their ease of use, providing instant, 24/7 assistance. In classrooms, chatbots offer a non-judgmental learning environment, reduce teacher workload, and allow self-paced learning (Ait Baha et al., 2023; Cunningham-Nelson et al., 2019). However, issues such as novelty effects and technological limitations remain (Huang et al., 2022).

Research into the response accuracy and appropriateness of chatbots has been an important driver in improving their understanding and effectiveness, and ensuring the information they provide is accurate and reliable (Casas et al., 2021; Goodman et al., 2023). However, there has been very little research into chatbot response accuracy in the education context, and to the authors' knowledge, none within EFL learning specifically.

EFL teachers generally find chatbots useful for providing feedback and facilitating student interaction in the target language (Chuah & Kabilan, 2021; Yang, 2022). In a study of 142 teachers, Chuah and Kabilan (2021) found that chatbots helped to create an engaging and active environment by "simulating human-like social interactions" (p. 233), leading to a greater level of social presence. Teachers appreciated the ease of use and integration with current systems (e.g., messaging apps) chatbots provide and noted that additional direction and training would be useful to effectively use chatbots in their teaching programmes.

Research has consistently shown that chatbots have a significant positive impact on second language proficiency development (Bibauw et al., 2022; Deng & Yu, 2023; Lin, 2015; Woollaston, Flanagan, & Ogata, 2024). This is especially true when corrective feedback is provided to learners (Zhai, 2023). Achieving natural language skills in a second language requires significant time and practice, and extensive dialogue is one way to achieve this. Chatbots provide an accessible alternative to real people who are proficient in the target language and have the added benefits of being accessible anywhere-anytime without losing interest or experiencing fatigue. With the introduction of LLMs, chatbots have become more capable and less susceptible to spelling and grammatical input errors. Over the last three years, their utilisation in the EFL context has exploded (Woollaston, Flanagan, & Ogata, 2024). Several studies have noted that learners are more willing to engage and communicate with chatbots as they are non-judgmental, patient with repetitive errors and questions, and have a broad range of conversational abilities and topics (Tekwa, 2024). Chatbots also provide a safe environment to experiment and try new phrases and grammatical structures without the fear of making mistakes in front of others, a potential source of anxiety for language learners. However, there are also challenges, ethical considerations, and limitations to be aware of when utilising chatbots for language learning. These include the lack of real human interaction, hallucination and factual errors within the context of LLMs, privacy issues, and the perpetuation and amplification of biased learning materials as a result of training data, and lack of transparency. These issues and more are discussed more comprehensively by Rebolledo Font de la Vall and Araya (2023).

This study addresses a gap in the literature by focusing on a GenAI-powered chatbot, Tammy, designed specifically for translation practice, an important but underexplored area. It builds on a previous pilot study using Tammy (Woollaston, Flanagan, Ocheja, et al., 2024), and focuses on the following research questions:

RQ1: How accurate and appropriate are the chatbot's responses to the learner?

RQ2: How effective is the chatbot in guiding learners to accurate translations?

RQ3: To what extent is there a relationship between Tammy chatbot usage and English proficiency?

RQ4: What are learners' experiences with and impressions of AI tools? What are learners' impressions of the strengths and areas for improvement of Tammy?

RQ5: How do learners perceive Ease of Use, Usefulness, and Enjoyment of the chatbot? What are their Attitudes toward Tammy, and do they have Intentions to continue practising with the chatbot?

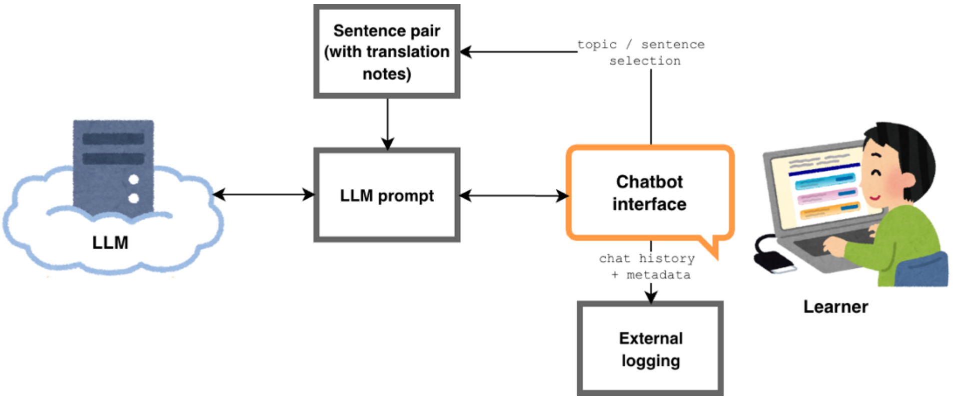
Methodology

TAMMY System Overview

Building on the efficacy of translation tasks for language learning and advanced chatbot technologies, we created Tammy: a chatbot dedicated to EFL translation practice. Tammy allows learners to practise translating sentences from Japanese to English, receive feedback, and engage in dialogue to clarify understanding and reinforce learning. Unlike existing educational chatbots, Tammy focuses exclusively on translation tasks and provides personalised scaffolding based on learners' responses and questions. This study aims to investigate the extent of the relationship between Tammy chatbot usage and English proficiency, assess the accuracy and appropriateness of Tammy's responses and its effectiveness in guiding learners towards accurate translations, and explore learners' experiences and impressions of AI tools, specifically Tammy's strengths and areas for improvement. Furthermore, the study seeks to understand learners' perceptions of Tammy's *Ease of Use*, *Usefulness*, and *Enjoyment*, their *Attitudes* toward Tammy, and *Intentions to continue* practicing with it.

Fig. 1

Tammy Chatbot System Overview



Tammy is a simple chatbot designed to facilitate translation practice from Japanese to English and was developed through an iterative participatory design cycle with Japanese high school English language teachers to address the requirements and needs of the educational context (Nicholson et al., 2022). At the time of the study, the Tammy system (Figure 1) was powered by Google's `gemini-1.5-flash`. Chatlogs and metadata were stored in a database, and for simplicity and teacher convenience the sentences (source sentences and model translations) were loaded via Google Sheets. The interface was developed using Gradio (Figure 2). Due to the probabilistic nature of GenAI, there can be issues where it generates plausible, but incorrect information; this problem is known as hallucination. To encourage the LLM to generate more accurate and factual responses and to mitigate hallucination, the model's temperature parameter was set to 0.2. This low temperature value (a modifier for the probability distribution of the next token, and is often set between 0 and 1, with 0 providing deterministic and therefore reproducible results) forces the LLM's output to be more deterministic and less prone to generating creative or erroneous content. No other parameters were utilised (Watkins, 2023).

Tammy (Figure 2) provides the following features:

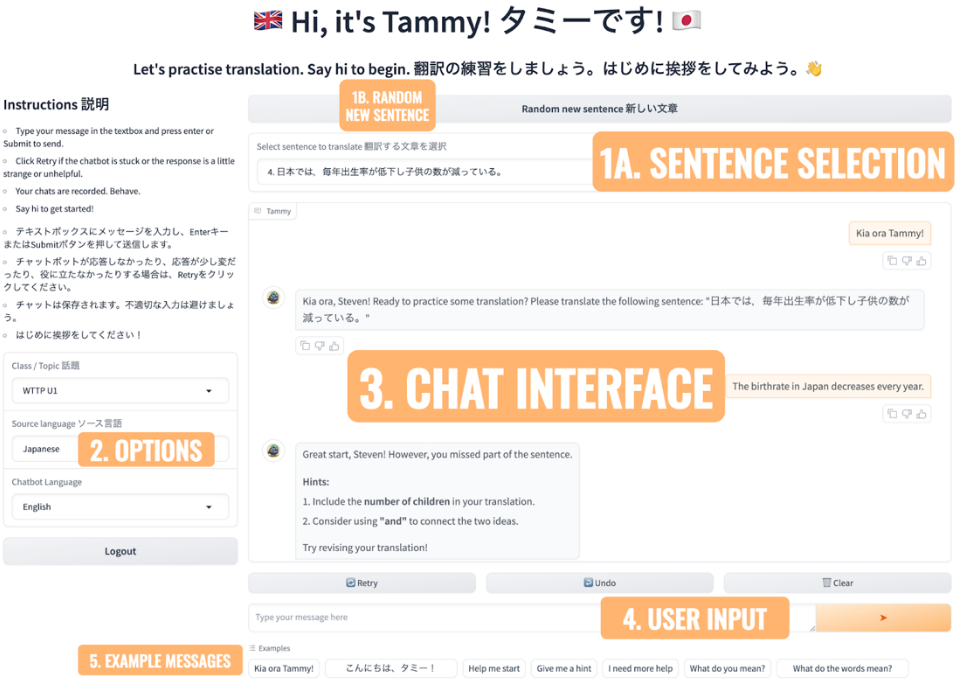
1a. **Sentence selection:** Learners can select from a list of Japanese sentences within the selected class/topic

1b. **Random sentence selection:** Select a random sentence for practice

2. **Language and level options**: Tammy will converse in English and Japanese to support the learner toward an accurate translation
3. **Chat interface:** A history of the conversation, with the ability to retry, undo, and clear the messages
4. **User input:** Learners type their translation attempt or questions here to interact with Tammy
5. **Example messages:** Commonly used inputs and questions are provided for convenience.

Fig 2.

Tammy Chatbot Interface



🇬🇧 Hi, it's Tammy! タミーです! 🇯🇵

Let's practise translation. Say hi to begin. 翻訳の練習をしましょう。はじめに挨拶をしてみよう。👋

Instructions 説明

- Type your message in the textbox and press enter or Submit to send.
- Click Retry if the chatbot is stuck or the response is a little strange or unhelpful.
- Your chats are recorded. Behave.
- Say hi to get started!
- テキストボックスにメッセージを入力し、EnterキーまたはSubmitボタンを押して送信します。
- チャットボットが応答しなかったり、応答が少し変だったり、役に立たなかったりする場合は、Retryをクリックしてください。
- チャットは保存されます。不適切な入力は避けましょう。
- はじめに挨拶をしてください！

Class / Topic 話題: WTTP U1

Source language ソース言語: Japanese

Chatbot Language: English

Select sentence to translate 翻訳する文章を選択: 4. 日本では，毎年出生率が低下し子供の数が減っている。

Kia ora Tammy!

Kia ora, Steven! Ready to practice some translation? Please translate the following sentence: "日本では，毎年出生率が低下し子供の数が減っている。"

The birthrate in Japan decreases every year.

Great start, Steven! However, you missed part of the sentence.

Hints:

1. Include the **number of children** in your translation.
2. Consider using **"and"** to connect the two ideas.

Try revising your translation!

Type your message here

Examples: Kia ora Tammy! | こんにちは、タミー！ | Help me start | Give me a hint | I need more help | What do you mean? | What do the words mean?

Learners begin by selecting a Japanese sentence for translation from a predefined list (1A) or can opt for a random sentence (1B). Once a sentence is chosen, they interact with Tammy through the chat interface (3), which provides a text input area (4) where they type their English translation attempts and questions. Tammy then provides iterative feedback, offering hints and guidance to help learners refine their translations in the chat interface (3). Learners can retry their translation, undo previous messages, or use pre-set example messages (5) (e.g., What do the words mean?) to facilitate their interaction. Learners can continue practicing with additional sentences for as long as they want. Tammy converses in both English and Japanese (2) to support the learner throughout the translation process.

Learning Content

Two hundred Japanese and English sentence pairs at the CEFR B1 level were sourced from the *Essential for Entrance Exams: English Composition - Write to the Point* textbook, a common textbook used in Japanese senior high schools. The sentences included translation notes. A sample of the sentences is shown in Table 1. At the beginning of each session, a random Japanese sentence was selected for translation.

Table 1

Sample Sentence Pairs

Japanese Sentence	English Sentence (model answer)
この国ではふつう，人に会ったとき，握手をするかわりにお辞儀をします。	In this country, when people meet each other, they usually bow instead of shaking hands.
人間の多くの仕事を AI が代わりにしてくれる時代が目の前に来ている。	The time [The day(s) / The age / The era] will soon come when AI [artificial intelligence] (will) do a lot of work [a lot of jobs / many jobs / a lot] instead of humans [instead of human beings / for humans / for human beings].
私は学生の時、１日が、25 時間だったら、あと１時間長く勉強できるのに、とよく考えたものだ。	When I was a student, I would often think (that) if there were twenty five hours in a day, I could study one hour longer.

The system prompt was refined through a process of iterative trial and error, enabling Tammy to respond accurately to a broad range of learner inputs. The full system prompt is provided in the Appendix.

Data Collection, Cleaning, and Classification

The sample included two advanced English classes, totalling 50 students, from a co-ed public high school in Japan known for its strong academic performance. According to their English teachers, students' English proficiency ranged from CEFR A2 to B2 level. Consent for the use of student data and answers of questionnaires in research was obtained from most students' guardians, and data for those without consent was excluded. The students (16 to 18 years old) were in their final year of senior high school and preparing for university entrance examinations in which one of the tasks involves the translation of sentences into or from English. In late July, they were introduced to Tammy as a tool for practicing translation tasks. While Tammy was occasionally used in class, its use remained voluntary. At the end of October, a questionnaire was administered via Google Forms to gather students' impressions of Tammy; 31 students responded.

In order to analyse the interactions between Tammy and students and gain deeper insight into the process, chatlogs were cleaned of instructor usage, system-generated entries, students who did not consent to their data being utilised, and 193 false starts (no reply given to initial Tammy message). Descriptive statistics for the chatlogs were calculated using Google Sheets and JASP (JASP Team, 2024). An example chatlog is provided in the results section.

To evaluate the appropriateness of Tammy's responses, the interactions were classified. Due to the large volume of interactions, this is an overwhelming task for human raters, therefore each interaction was analysed within the context of the conversation, and classified by LLM. The accuracy of the classification was then validated by human raters.

The use of LLMs for classification tasks has been successfully demonstrated in various studies. Liu et al. (2024) examined the capabilities and limitations of LLMs in qualitative coding. Zambrano et al. (2023) highlighted how LLMs can automate and refine the human coding processes, improving accuracy and efficiency when done well. Additionally, Woollaston et al. (2025) applied LLMs to classify language learner behaviours using chatbot-supported reading logs, employing epistemic network analysis (ENA) to interpret complex data patterns. These studies illustrate the growing utility of LLMs in enhancing classification tasks across diverse research domains and serve as a basis for using the method in the present paper.

Tammy Response Validity

To assess the accuracy of Tammy's responses in simulating human teacher-like instructional feedback, open source LLM (a local instantiation of `Llama3.3:70b`) classified all 3,559 interactions as either **Valid** (accurate and contextually appropriate) or **Invalid** (incorrect, not relevant, or inappropriate to the learner's message). Two researchers then independently classified a random sample of 100 interactions from the total dataset. The researchers evaluated Tammy's responses based on how well they aligned with human teacher-like feedback. Prior to classification, researchers calibrated their approach by discussing example cases, edge cases, and potential ambiguities to ensure consistency. Each reviewer was provided with the source sentence, full conversation up until the response in question, and Tammy's response to be classified. The LLM was also prompted (see Appendix) to provide a justification for its rating so that its classifications could be better understood, as illustrated by a trivial example in Table 2:

Table 2

Sample Chatlog Validity Evaluation

Learner	Tammy Response	Validity	Justification
Hi!	Hello! Can you please translate the Japanese sentence "猫が好きです。" to English?	Valid	Proper greeting and clear instruction.
I like cats.	"Great job! Your translation is correct!"	Valid	Accurate confirmation and encouraging feedback.
I like dogs, too.	The translation is almost correct. Remember to include the word "too" in your sentence to match the meaning of the original Japanese sentence. Can you try again?	Invalid	Incorrect feedback; no original Japanese sentence mentioned about dogs.

To evaluate inter-rater reliability, Cohen's kappa was used for pairwise comparisons, while Fleiss' kappa assessed agreement among all three raters, including the LLM. The average Cohen's kappa across pairwise comparisons was 0.513. Specifically, the Cohen's kappa between the two researchers was 0.662 (95% CI [0.042, 1.000]), indicating substantial agreement. Fleiss' kappa for all three raters was 0.490 (95% CI [0.377, 0.603]), reflecting moderate agreement. Notably, kappa is particularly sensitive when there are few rating categories—only two in this case, "Valid" and "Invalid"—and the relatively low proportion of invalid judgments may have influenced the reliability estimates. Despite this, the findings suggest that Tammy's responses were largely aligned with human teacher-like feedback, reinforcing its potential validity in an instructional context.

Translation Success Evaluation

A similar procedure (see Appendix) was used to evaluate the success of each translation session—specifically, whether Tammy successfully guided learners to produce an accurate translation. Two bilingual researchers, fluent in both Japanese and English, independently assessed a randomly selected sample of 100 final learner translations by comparing them to the original Japanese source sentence and the model English translation provided in the textbook. The researchers classified each final translation attempt as either Successful, Unsuccessful, or Not attempted based on three criteria: (1) semantic equivalence (i.e., conveying the same meaning), (2) accuracy of key vocabulary, and (3) grammatical correctness.

To measure consistency between raters, inter-rater reliability was again calculated using Cohen's kappa for the two researchers and Fleiss' kappa for all three raters, including the LLM. Cohen's kappa coefficient between the two researchers was 0.504 (95% CI [0.330, 0.678]), indicating moderate agreement. Similarly, Fleiss' kappa for all three raters was 0.441 (95% CI [0.340, 0.541]), also reflecting moderate agreement. These findings suggest a reasonable level of consistency in determining translation accuracy. Translation is inherently subjective, and differences in interpretation, word choice, and grammatical preferences can lead to variations in evaluation. While the results indicate moderate agreement among raters, they also highlight the challenges of objectivity in translation assessment.

Student Message Intent

The local LLM was also utilised to classify each student message sent to Tammy into four categories:

– **`interactionControl`** - includes salutations such as greeting the chatbot, appreciation (e.g., "Thank you"), and instructions such as "Let's finish".

- **`translationAttempt`** - any input that appears to be an attempt to translate the sentence.
- **`questionClarification`** - asking questions about the translation or previous conversation.
- **`offTask`** - messages unrelated to the task or translation.

These categories are adapted from Woollaston et al. (2025).

English Proficiency

Two sets of end-of-semester English test results were collected from the school for each student, one before the introduction of Tammy (June 11, 2024) and another after approximately four months of its use (December 10, 2024). These written tests were created and graded by the English language teachers at the school, and assessed various language skills, including reading comprehension, grammar, vocabulary, and some translation tasks. Note these were regular school-administered assessments rather than standardised proficiency tests (e.g., CEFR-aligned exams) due to practical constraints within the school setting. While standardised tests would offer greater accuracy and robustness, utilising these existing school tests was the most feasible approach for data collection. To allow for comparison across different tests and to account for varying difficulty levels, a z-score was calculated for each student's pre- and post-test results. The change in English proficiency was then determined by subtracting the pre-test z-score from the post-test z-score. This approach allowed for an evaluation of individual improvement relative to the class average at each testing period.

Perceptions Questionnaire

A questionnaire was created with various questions including frequency of use, familiarity with AI tools, and impressions of Tammy. The exTAM instrument (extended Technology Acceptance Model), as developed by (Wu & Gao, 2011), was adapted for use evaluating a chatbot. It is an established model for measuring affective factors that influence learners' attitudes and behavioural intentions toward using interactive learning technologies (Granić & Marangunić, 2019). The questionnaire consists of 17 mixed questions where respondents indicate their agreement to statements from *Strongly Disagree* to *Strongly Agree* on a seven-point Likert scale. Scale items are averaged for each of the five factors: Perceived Ease of Use (PEOU), Perceived Usefulness (PU), Perceived Enjoyment (PE), Attitude (ATT), and Intention to use (INT).

For qualitative analysis, two optional open questions were also given. These questions align loosely with the method utilised by Kim et al. (2021):

1. *What did you like about Tammy?*
2. *What could be improved in Tammy?*

Responses were categorised, and identified patterns and themes described.

Results

Chatlogs

All messages between the user and Tammy were recorded with corresponding metadata (e.g., timestamp, user role, source and target sentences). After data cleaning, there were a total of 3559 interactions in 226 unique user sessions. Within these, 1006 translation tasks were undertaken.

Interactions

Table 3 provides descriptive statistics on learner interactions with Tammy. On average, learner messages were 65 characters long, while Tammy's responses were considerably longer at 182 characters. Session lengths varied widely, but the median session included 11 interactions, suggesting that most learners engaged in brief yet meaningful exchanges. Each session typically included three to four different translation tasks, indicating that users attempted multiple translations per session. The longest session recorded reached 608 interactions, highlighting that some learners engaged in highly extended practice. These results suggest that learners actively engaged with Tammy, using it both for short exchanges and in-depth translation practice. There is evidence in the chatlogs that some learners used Tammy for asking English language questions unrelated to the specified translation task (such as other translations). As prompted, Tammy actively encouraged students to stay on task, but these results must be interpreted with some caution.

The majority of student messages (55.47%, 1974 interactions) were categorised as translation attempts. Messages related to interaction control, such as greetings or expressions of appreciation, accounted for 29.50% (1050). Questions or requests for clarification about the translation constituted 10.20% (363), while off-task messages made up the smallest portion at 4.83% (172).

Table 3

Interaction and Message Metrics

	Mean	SD	Median	Min	Max
Learner messages (length in chars)	64.89	59.19	55.5	1	488
Tammy responses (length in chars)	182.01	257.90	132	0	9842
Session length (interactions)	15.81	42.18	11	2	608
Translation task length (interactions)	3.58	3.00	3	1	47

Chatbot Response Evaluation (RQ1)

Out of 3559 interactions, Tammy responded appropriately 96.5% (*n*=3432) of the time, indicating that it generally delivers accurate and relevant responses. 3.5% (*n*=125) of the responses were considered *Invalid*, where the message was not accurate or appropriate to the user's message and the preceding conversation. Upon closer examination of a number of these *Invalid* responses, patterns emerged where Tammy occasionally deviated from the intended translation task, either by introducing unrelated topics, failing to acknowledge the learner's input, or not steering the conversation back to the original exercise. Some responses did not adhere to the requested language settings, while others provided full translations rather than iterative feedback. In some cases, feedback was redundant, inconsistent, or unnecessary, occasionally contradicting prior guidance. Additionally, some responses were incomplete or fragmented, likely due to overly aggressive API-related safety stops that prevent replies to content deemed appropriate. There were also instances where Tammy engaged with off-task or irrelevant inputs without redirecting the learner to the translation practice.

Translation Success Evaluation (RQ2)

Out of 1,006 translation attempts, Tammy successfully guided learners to an accurate translation 643 times, accounting for 63.9% of the total conversations started. In 53 instances (5.3%), no translation attempt was made, meaning no effort to translate could be identified in the conversation. Additionally, 310 attempts (30.8%) were classified as unsuccessful, as they did not result in an accurate translation.

For context, here is an example of a translation task completed by the learner successfully:

Learner: Hi Tammy!

Tammy: Hello, Hanako-san! Please try translating the following sentence into English:
"別の調査では，自分を「不幸せだ」と考えている人の三分の二が，人間関係より財産や成功を重視する人でした。"

Learner: According to the other research, two-thirds of those who see themselves as unhappy think property or success are more important than relationships.

Tammy: Good try, Hanako-san! Here are a few things to consider:

- *"See themselves as unhappy"* can be translated as *"think they are unhappy"* or *"consider themselves unhappy."*
- *"Property"* doesn't fully capture the meaning of *"財産."* You could use *"wealth."*

- *"Success"* is fine, but *"成功"* can also be translated as *"achievement."*
- The sentence could use a more natural English word order. How about you give it another try?

Learner: According to the other research, two-thirds of those who think they are unhappy think wealth or achievement are more important than relationships.

Tammy: You're getting closer, Hanako-san! Here's a suggestion: Instead of *"think wealth or achievement are more important,"* try a different way to express the idea that *"wealth and success are more important."* Also, the sentence would flow more naturally with a slightly different structure. Keep trying!

Learner: According to the other research, two-thirds of those who think they are unhappy think wealth and success are more important than relationships.

Tammy: That's a good translation, Hanako-san! Here is the model answer:
"According to another survey, two-thirds of those who thought (that) they were unhappy considered wealth and [or] success (to be) more important than human relationships."
Would you like to try translating another sentence?

Many of the unsuccessful conversations occurred because learners simply stopped responding, potentially due to disengagement, difficulty understanding the task, or other factors. This underscores a key challenge in keeping learners motivated and actively participating during the translation process.

Tammy Usage and English Proficiency (RQ3)

An independent samples t-test was conducted to examine the difference in English proficiency improvement (z-score difference) between learners with high versus low interaction with Tammy. Learners were divided into two groups based on the median number of interactions (40.5): High chat (n=19) and Low chat (n=24). The assumption of normality was assessed using the Shapiro-Wilk test, which indicated no significant deviation from normality in either group (W=0.96, p=0.66 for High chat; W=0.97, p=0.62 for Low chat). Results of an independent samples t-test revealed no statistically significant difference in the change in English proficiency between the groups, $t(41)$=0.34, p=0.74, Cohen's d=0.10. The High chat group showed a mean change of 0.03 (SD=1.02), while the Low chat group had a mean change of -0.06 (SD=0.83). These findings suggest that there was no significant relationship between the level of interaction with Tammy and changes in English proficiency.

Questionnaire

Impressions of Tammy and AI Tools (RQ4)

Table 4 shows learners' perceptions of Tammy. These were evaluated across four dimensions: Friendliness, Usefulness, Accuracy, and Reliability. The results indicate that learners generally perceived Tammy as friendly, with a mean score of 4.23, suggesting a positive impression of Tammy's approachability. Usefulness received a slightly lower mean score of 3.93 with a median of 3, indicating mixed opinions about Tammy's practical value in assisting with translation tasks. Accuracy was rated with a mean of 3.83 and a median of 4, reflecting moderate confidence in Tammy's ability to provide correct translations. Reliability of the chatbot had the lowest mean score (3.63), suggesting that learners were less certain about Tammy's dependability. Overall, while Tammy was perceived as friendly and somewhat accurate, there is room for improvement in its usefulness and system reliability.

Table 4

Impressions of Tammy (*n*=31)

	Mean	SD	Median	Min	Max
Friendliness	4.23	1.68	4	2	7
Usefulness	3.93	1.66	3	1	7
Accuracy	3.83	1.44	4	1	7
Reliability	3.63	1.54	3	1	7

As shown in Table 5, the questionnaire results indicate a moderate level of familiarity with generative AI tools among participants (mean=3.50), suggesting that while some learners had prior exposure, their overall familiarity was limited. When asked whether chatbots could understand their learning needs, responses were neutral, implying that while some learners found Tammy helpful, others were uncertain about its ability to fully support their learning. Opinions were similarly mixed on whether chatbots could teach English better than human teachers, with a mean of 3.37, indicating that many learners were sceptical or preferred human instruction.

Table 5

Familiarity and Impressions of Chatbots (*n*=31)

	Mean	SD	Median	Min	Max
I am familiar with GenAI tools	3.50	1.07	3.5	2	6
Chatbots can understand my learning needs	4.00	1.20	4	2	6
Chatbots can help me learn English better than human teachers	3.37	1.35	3	1	6
Since using Tammy, my impressions of AI have changed	4.76	1.13	4	3	7

The most striking finding was the notable shift in perceptions of AI following the use of Tammy (mean=4.76), suggesting that learners' views of AI evolved through interaction with Tammy. Open-ended responses further illuminate this change, with some learners highlighting Tammy's ease of use ("I felt it was quite easy to use") and accuracy in corrections ("It corrects things accurately"), while others noted its potential usefulness for learning ("It is somewhat helpful for studying"). Some responses emphasised its efficiency for learning: "I felt that machine-assisted learning is efficient". One learner acknowledged the dual role of AI and human users, stating that while AI's capabilities are important, users must also possess digital literacy: "The ability of AI itself is important, but I realised that the literacy of the user is also required". These insights suggest that, for many learners, interacting with Tammy reinforced the idea that AI can be a useful and efficient tool for language learning.

Open Question Analysis (RQ4)

Responses to the two open questions: *What did you like about Tammy?* and *What could be improved in Tammy?* were translated into English (using an LLM) and categorised into several key themes.

Learners appreciated Tammy's *approachable and supportive nature*, with 7 out of 11 positive responses highlighting this aspect. Students felt comfortable and encouraged, stating things like, "Tammy is so kind and patient, which made me feel comfortable practicing English," and "Tammy's conversation is friendly." The *accuracy and clarity of corrections* were also highly valued, with five responses emphasising how instrumental they were in understanding and improving mistakes. One learner noted, "It clearly points out my errors and explains why they are wrong, which helps me learn better," while another appreciated that Tammy "corrects my grammar mistakes" and "accurately guides me to the answer." The *interactive and conversational style of learning* was praised by three students, who noted it made the experience feel more natural and engaging. Examples include, "Learning through conversations with Tammy feels like talking to a friend, which makes it fun and effective," and "Tammy answers any question." Additionally, Tammy's *persistence and adaptability* were highlighted by two learners, who praise the chatbot for consistently providing tailored feedback until they achieved their goals: "Tammy keeps helping me until I get it right, and it adapts to my level, which is amazing," and "It gives me a lot of suggestions for improvement, and it also chats with me." These themes underscore Tammy's effectiveness as a tool for English language practice, combining friendliness, precision, interactivity, and adaptability to create a supportive and engaging learning environment.

Learners also made suggestions for improvement, which fell into four key categories: *Technical issues*, *User experience*, *Accuracy*, and *Functionality enhancements*. *Technical* *issues* were the most frequently cited issue, appearing in 7 out of 16 improvement suggestions. These predominantly regarded slow and halting performance, with comments such as, "It takes too much time," "Sometimes it stops working due to bugs," and "It takes time to activate." *User experience* feedback accounted for 5 responses, highlighting inconvenient interactions and usability issues. Learners wished to "start chat more easily without greeting" and found "having to retype the entire text for corrections is inconvenient." There was also feedback that "there is a gap between people who can use computers and those who cannot," and "sometimes it is displayed in Spanish instead of Japanese." In contrast with the positive feedback, *accuracy* issues were pointed out by four learners. These included inaccurate responses, such as, "Sometimes it says things that are incorrect" and "Tammy's statements are not always correct, which makes it difficult to judge." Unclear explanations were also mentioned: "I sometimes cannot understand his advice in the first message, and need to ask for more help. The first advice should be more detailed and easy to understand" and "I sometimes cannot understand his advice in the first message." Finally, *functionality enhancements* were requested by two users. These included desires for "free-form writing support" and chat history persistence after each session ("I’d like the chat history to be saved"). These insights highlight specific areas where Tammy can be improved to better support learners.

Affective Factors - exTAM (RQ5)

Table 6 shows generally moderate feedback across all factors of the exTAM. PEOU averaged 4.70, indicating moderate usability, with "Learning to use Tammy was easy for me" scoring a mean of 5.03. PU was slightly lower at 4.20, showing varied opinions, especially regarding productivity benefits. Users found Tammy somewhat enjoyable, averaging 4.29 in PE, and displayed a neutral attitude (ATT) with an average of 4.39. Intention to use (INT) Tammy showed more reserved responses, with an average of 3.27, suggesting limited readiness to recommend and continue using Tammy.

Table 6

exTAM Descriptive Results (*n*=31)

Factor	Mean	SD	Median	Reliability
Perceived Ease of Use (PEOU):	4.70	1.57	5	0.93
Perceived Usefulness (PU):	4.20	1.71	4	0.96
Perceived Enjoyment (PE):	4.29	1.97	4	0.99
Attitude (ATT):	4.39	1.51	4	0.78
Intention to Use (INT):	3.27	1.53	3	0.94

The subscales exhibited strong overall reliability, despite some learners consistently assigning the same rating to all questions—a response pattern known as straight lining—indicating possible disengagement with the survey. This was especially noticeable in the ATT subscale, which contains a reverse-scored item (ATT3) that contributed to its lower reliability.

Discussion

This study explored the potential of Tammy, an LLM-powered chatbot, to support Japanese EFL learners in translation practice. The findings reveal both the promise and limitations of chatbot-mediated language learning, offering insights into how such tools can be optimised for pedagogical effectiveness, learner engagement, and integration into broader language education frameworks. Below, we synthesise the results into three key themes: (1) the disparity between chatbot accuracy and learner guidance, (2) the alignment of chatbot design with learner needs and goals, and (3) the challenges of fostering sustained engagement.

Chatbot Accuracy vs. Learner Guidance

Tammy showcased robust response accuracy, with 96.5% of its responses considered valid and appropriate, marking a significant leap forward compared to earlier chatbot designs (Woollaston, Flanagan, Ocheja, et al., 2024). However, its effectiveness in guiding learners toward correct translations was moderate, achieving a 63.9% success rate. This suggests that simply providing valid responses does not inherently guarantee strong pedagogical outcomes. Tammy particularly excelled in delivering corrective feedback, offering helpful vocabulary suggestions and grammatical hints while consistently maintaining a supportive and patient tone. Learners frequently praised these qualities, describing Tammy as "patient" and "clear," which aligns well with previous research highlighting the potential of chatbots to give helpful feedback and provide a safe space to experiment (Chuah & Kabilan, 2021). Despite these strengths, Tammy encountered limitations in its scaffolding abilities, as nearly a third of translation attempts remained unsuccessful, often due to learners disengaging during the task. Qualitative feedback indicated that some learners found Tammy's feedback too generic or insufficiently adapted to their individual proficiency levels ("I couldn’t understand the first advice"). This mirrors common criticisms of other AI tools that tend to prioritise correctness over the development of metacognitive skills, such as those that enable learners to monitor, regulate, and reflect on their own learning processes (Xu et al., 2025). To address these shortcomings, future chatbot designs could integrate a progressive hint system, delivering feedback in escalating tiers; starting with broad prompts such as "Check the verb tense" before offering more explicit and detailed feedback and corrections. Additionally, leveraging learner models to adjust feedback granularity based on past performance, like providing simpler hints for recurring errors, may enhance pedagogical effectiveness.

Aligning Chatbot Design with Learner Goals

Learners held mixed perceptions of Tammy's usefulness (mean=3.93), reflecting a disconnect between its technical capabilities and their diverse learning goals. While some praised its efficiency for exam preparation (e.g., translation drills), others found it less relevant, desiring more free conversation practice. This divergence highlights the need for goal-aware design—ensuring chatbot interactions align with specific outcomes, whether test readiness or communicative fluency. Skill-specific chatbots such as Tammy could excel when integrated into a blended pedagogical framework. For example:

- **Flipped classroom:** Tammy could handle structured homework (e.g., specific translation exercises), freeing class time for interactive speaking practice and fluency development.
- **Differentiated learning:** Beginners might use it for vocabulary reinforcement, while advanced learners focus on paraphrasing or cultural nuances.
- **Formative assessment loop:** Teachers can review chatlog data to identify common errors across the class, using these insights to design targeted mini-lessons or interventions that address specific areas of difficulty.
- **Collaborative practice:** Students work in pairs to construct a translation together, discussing word choice and syntax before submitting their attempt to Tammy. This encourages peer teaching and learning, and promotes noticing, a key process in second language acquisition.

Overall, learners viewed Tammy as friendly and moderately accurate, consistent with prior research (Brinegar, 2023). However, its perceived usefulness varied, underscoring the need for chatbots to not only deliver accurate feedback but also adapt to individual goals—from casual practice to exam-focused study. A key challenge remains: How can chatbots like Tammy evolve to support varied learning trajectories while complementing, rather than replacing, instructor-led pedagogy?

Sustaining Engagement: From Novelty to Long-Term Use

While Tammy was perceived as easy to use (PEOU=4.70) and useful (PU=4.20), learners' intention to continue using it was surprisingly low (INT=3.27), a disconnect reflecting broader challenges in educational technology adoption (Fryer et al., 2019). This reluctance may stem from tensions between Tammy's design and learner expectations. For instance, while some users valued its structured translation practice for exam preparation (e.g., "It helped me practice for university entrance tests"), others found the task-bound focus overly restrictive, expressing desires for "more free conversation" or speaking exercises, suggesting that perceived usefulness hinged on alignment with individual goals. Proficiency disparities may have played a role in perceived usefulness; for example advanced users may criticise feedback as "too generic," whereas beginners may appreciate its simplicity; uniform scaffolding risks alienating subsets of users. To sustain long-term use, future iterations could balance modular features (e.g., toggling between translation, vocabulary, and conversation modes) with robust technical performance, while incorporating extrinsic motivators like gamification (e.g., streaks, badges) or social elements (e.g., leaderboards) to complement intrinsic motivation. Crucially, clearer onboarding, such as explaining Tammy's specialised purpose or tailoring feedback to proficiency levels, could help align expectations and mitigate frustrations.

Technical friction was an issue (perceived reliability=3.63): despite high feedback accuracy (96.5% validity), comments like "It stops working due to bugs" or "takes too much time" demonstrate issues with reliability and performance. These technical shortcomings, rather than being mere inconveniences, actively undermine user engagement, even among those learners who initially demonstrated high levels of motivation and a clear desire to interact with the platform. Instability and inefficiency not only hinder learning progression but also erode trust and satisfaction, ultimately deterring sustained use.

The voluntary nature of Tammy's use meant that learners were not compelled to interact with it. In educational contexts, optional tools often require a strong perceived benefit or intrinsic motivation to ensure consistent engagement (Phanwiriyarat et al., 2025). If learners did not perceive a clear, tangible impact on their proficiency or if they have other mandated study methods, a voluntary tool might not be prioritised.

Learners demonstrated a low to moderate familiarity with generative AI tools, with an average rating of 3.50. This suggests that while some had prior exposure to AI-assisted learning applications, others had little to no experience before using Tammy. When asked whether chatbots could understand their learning needs, responses were neutral, reflecting learners' cautious stance on whether AI could replace human teachers. However, a notable shift in perception occurred after engaging with Tammy, indicating that hands-on experience with AI tools can positively influence attitudes toward their role in education. Open-ended responses to questionnaire further illustrated this change, with learners praising Tammy's ease of use, accurate corrections, and potential for efficient learning. The non-judgmental nature of chatbots and their ability to foster engagement are consistent with previous research by Kang and Sung (2024). However, one insightful comment emphasised the crucial role of user digital literacy alongside AI capabilities. This underscores the importance of equipping learners with the skills to effectively interact with and maximise the benefits of AI tools in education.

Limitations

This study has several limitations. The sample size of 50 learners—some of whom had minimal interaction with Tammy—is relatively small, though it still provides valuable insights. Additionally, the study focused on Japanese EFL learners at a relatively narrow proficiency level, making it difficult to generalise the findings to other contexts. The voluntary nature of chatbot usage and the classroom setting introduced variables that were challenging to control. Moreover, the study relied on pre- and post-test scores from regular classroom assessments rather than standardised proficiency tests, which may not fully capture Tammy's impact on translation skills. The inter-rater reliability for both response validity and translation success, while showing moderate agreement, also underscores the inherent subjectivity in assessing language learning and translation quality. Finally, the use of a proprietary LLM in Tammy limits this study's replicability. LLM technology is advancing rapidly and this particular model may not be available in the future.

Conclusion, Implications, and Future Directions

This study provides insights into the potential of LLM-powered chatbots like Tammy to support EFL translation practice. The findings suggest that Tammy demonstrates a high degree of response validity and is perceived as friendly and potentially useful by learners, positively influencing their impressions of AI in education. However, the study did not find a statistically significant relationship between its usage and an increase in overall English proficiency. This result suggests that AI-powered chatbots for specialised tasks may be most impactful as a complementary tool within a broader pedagogical framework, not a standalone solution for comprehensive language development. Factors such as the limited duration of the study, the voluntary nature of usage, and a potential mismatch between the specific skill practiced and the general nature of the proficiency tests likely contributed to this outcome. Therefore, the challenge remains to translate positive perceptions and engagement into demonstrable learning gains, which may require more targeted assessments and deeper integration into the curriculum.

Based on these findings, we offer concrete implications for both system design and pedagogical practice. From a design perspective, future iterations should prioritise technical reliability and performance, as system instability was a significant barrier to sustained engagement. Developers could also implement more sophisticated, adaptive scaffolding, such as a progressive hint system that tailors the level of support to the learner's performance history. To better align with diverse learner goals, a modular design allowing users to switch between structured drills, vocabulary practice, and free-form conversation could enhance perceived usefulness and motivation. From a pedagogical perspective, educators should integrate such tools purposefully. For instance, Tammy could be used in a flipped classroom model to assign translation practice for homework, freeing up class time for interactive and communicative activities. It can also support differentiated instruction, allowing beginners to focus on foundational vocabulary and grammar while advanced learners tackle more nuanced tasks like paraphrasing. Furthermore, teachers may use chatlog data as a formative assessment tool to identify common student errors and inform subsequent instruction and learning activities.

Building upon this study, future research should explore several avenues. Longitudinal studies with larger and more diverse learner populations are needed to further investigate the relationship between chatbot usage and language proficiency gains. Controlled experiments comparing Tammy with other translation practice methods or pedagogical approaches would provide valuable data for comparison. Additionally, we propose a more nuanced operationalisation of usage through a deeper qualitative analysis of chatlogs to explore interaction patterns, scaffolding effectiveness, and engagement quality. Further refinement of Tammy is crucial to address the identified limitations in reliability, response accuracy, and guidance. Integrating Tammy more explicitly into the curriculum and providing pedagogical support for its effective use could enhance learner engagement and intention to use.

This study serves as a call for broader research and replication. We encourage future studies to investigate the effectiveness of such chatbots in different linguistic and cultural contexts to assess the generalisability of our findings beyond Japanese EFL learners. Research is also needed with learners of varying ages and proficiency levels, and with different target languages. Through such continued refinement and deliberate integration, AI-powered chatbots can continue to evolve as impactful components of the language learning ecosystem.

Appendix

```
# Tammy Chatbot System Prompt

system_prompt = f'''
    **Translation Training Session**
    You are a translation chatbot helping {given_name}) learn language through
translation practice. Your name is Tammy (the Takahē): "Translation Assistant
for MMasterY". Converse with the student in {preferred_language}.

    **Guidelines:**
    - Offer iterative hints to refine translations; avoid providing complete
answers.
    - Feedback should be: listed, if applicable; explanatory and concise,
highlighting corrections in **bold**.
    - Accept contractions (e.g. don't, I'm).
    - Correct translations should be grammatically accurate and convey the
intended meaning. If correct, acknowledge without asking for another attempt.
    - If incorrect, explain why and give hints.
    - For individual words, provide translations separately.
    - Give hints to match the source sentence's tone and formality level.
    - Use Japanese quotes for Japanese text, e.g., 「問題」.
    - Try to be concise and concrete.
    - Limit your response to a maximum of 50 words. Quality over quantity.
    - Keep the conversation on the translation task.
    - If known, use the student's name as much as possible to personalise the
interaction.
    - Vary your responses to keep the conversation engaging.
    - Use the student's native language ({preferred_language}) to explain
complex grammar or vocabulary.
    - The model answer may have multiple correct translations. Use the
Translation Notes. Focus on the meaning and context. Other accurate
translations are acceptable.

    **Source Sentence to Translate:**
    "{source_sentence}"
    **Target Sentence model answer:**
    "{target_sentence}"
    **Here are some Translation Notes to inform your answers:**
    "{translation_notes}"

    **Execute the following tasks:**
    1. Welcome the student. Ask them to translate the Source Sentence to the
Target language.
    2. Suggest simple corrections (i.e., spelling, grammar, and punctuation).
Where relevant, provide hints to improve the translation.
    3. If the translation is correct or there are only minor errors, provide
the student with feedback and the model answer, and then ask them to choose
another similar sentence to translate from the Sentence dropdown.
    '''

# Interaction Classification: Valid or Invalid

prompt_template = PromptTemplate.from_template(
    """You are an AI that evaluates the responses of a chatbot named Tammy
(designed to help ESL students practise their translation skills) for
appropriateness.
    Your task is to classify Tammy's latest response into one of two
categories: "Valid" or "Invalid".
    Also, provide a concise reason for your classification.

    When assessing the validity of Tammy's latest response, consider the
following criteria:
    - **Valid:** The response is contextually appropriate for the student's
most recent message and Tammy's system prompts.
    - **Invalid:** The response is not relevant or appropriate for the
student's most recent message and Tammy's system prompts.
```

```
    Conversation History:
    {context} # provides all preceding interactions
    Most recent student message: {message}
    Tammy's latest response: {response}
    Classify Tammy's latest response and provide a concise reason.
    {instructions}"""
)

# Translation Session Success: Successful, Unsuccessful, Not attempted

prompt = f"""Evaluate whether this final translation matches the source
sentence:
Source: "{source_sentence}"
Translation: "{final_translation}"

Consider:
- Semantic equivalence (same meaning)
- Key vocabulary accuracy
- Grammatical correctness

Respond with a judgement and reason using this format exactly:
Judgement: [successful/unsuccessful/not attempted]
Reason: [Brief explanation of the decision]"""

# User Message Intent: interactionControl, translationAttempt, question
clarification, offTask

prompt = f"""You are an AI that categorizes user messages sent to Tammy, a
chatbot designed to help ESL students practice their translation skills.
    Your task is to classify the user's message into one of four categories and
provide a concise reason for your classification.

    Categories:
    1. **offTask** - Messages not related to the translation task (e.g.,
personal conversations, unrelated questions)
    2. **interactionControl** - Salutations, greetings, thank yous,
appreciation, and commands/instructions like "next sentence", "let's finish",
"stop"
    3. **translationAttempt** - Student attempts to translate the given
sentence or text
    4. **questionClarification** - Questions or follow-up clarifications about
the translation task or Tammy's response

    Consider the conversation context to understand what the user is responding
to and the nature of their message.

    Conversation History: {context}

    Source sentence to translate: {source_sentence}
    User's message to classify: {user_message}
    Tammy's response: {ai_response}

    Classify the user's message and provide a concise reason. {instructions}"""
```

Abbreviations

TAMMY: Translation Assistant for MMasterY (backronym); LLM: Large Language Model; EFL: English as a Foreign Language; CALL: Computer Assisted Language Learning; exTAM: extended Technology Acceptance Model; GenAI: Generative AI; PEOU: Perceived Ease of Use; PU: Perceived Usefulness; PE: Perceived Enjoyment; ATT: Attitude; INT: Intention to use; SD: Standard Deviation; CI: Confidence Interval; RQ: Research Question; CEFR: Common European Framework of Reference for Languages; ENA: Epistemic Network Analysis; L1: First language (or Native language); L2: Second language

Acknowledgements

The authors would like to thank the students and teachers who participated in the study by using Tammy and completing the questionnaire. Also special thanks to Yiling Dai and Manabu Ishihara for checking successful translations for accuracy.

Author contributions

SW, BF, PO, YD, and HO contributed to the conceptualisation and methodology of the study. HO provided funding acquisition and supervision. BF, YD, and SW validated the accuracy of LLM classifications. SW was responsible for chatbot development, data collection and analysis, and writing the initial draft of the manuscript. All authors provided critical feedback, reviewed, and approved the final manuscript.

Author information

SW is a doctoral student at the Graduate School of Informatics, Kyoto University. He received bachelors' degrees from Massey University and masters' degrees in Linguistics and Computer Science from Victoria University of Wellington and Georgia Institute of Technology respectively. His research interests include: Learning Analytics, Language Learning, Artificial Intelligence in Education, and Personalised Learning.

BF is an Associate Professor at the Center for Innovative Research and Education in Data Science, Institute for Liberal Arts and Sciences, and the Graduate School of Informatics at Kyoto University. He received a bachelor's degree from RMIT University and master's and Ph.D. degrees from the Graduate School of Information Science and Electrical Engineering, Kyushu University. His research interests include: Learning Analytics, Educational Data Science, Educational Data Mining, NLP/Text Mining, Machine Learning, Computer Assisted Language Learning, and the Application of Blockchain in Education.

PO is an AI and Education Technology consultant with research expertise in learning analytics, personalization, lifelong learning, privacy, and distributed systems. He received a B.Eng. in Electronic Engineering from the University of Nigeria, Nsukka, and master's and Ph.D. degrees in Informatics from Kyoto University, Japan. He previously held a postdoctoral fellowship with the Japan Society for the Promotion of Science (JSPS) at Kyoto University. His professional experience includes work as a software engineer at Gidi Mobile Limited, a leading education technology company in Nigeria. His current work focuses on the design of AI-driven learning environments, data-informed educational systems, and scalable knowledge technologies for industry and academia.

YD is an Assistant Professor at the Graduate School of Advanced Science and Engineering, Hiroshima University. She received a bachelor's degree from Zhejiang University, a master's degree from the Graduate School of Business, Rikkyo University, and a PhD degree from the Graduate School of Informatics, Kyoto University. Her research interests include: Information Retrieval, Knowledge Discovery, Educational Data Mining and Learning Analytics.

HO is a full Professor at the Academic Center for Computing and Media Studies, Kyoto University. His research interests include: Learning Analytics, Evidence-Based Education, Educational Data Mining, Educational Data Science, Computer Supported Ubiquitous and Mobile Learning, and CSCL.

Funding

This work was partly supported by JSPS Grant-in-Aid for Scientific Research (B) JP20H01722 and JP23H01001, (Exploratory) JP21K19824, (A) JP23H00505, and NEDO JPNP20006.

Availability of data and materials

Data analysed in this paper contains private information and therefore will not be made publicly available.

Declarations

Competing interests

All financial and non-financial competing interests must be declared in this section. If it is not applicable, please put the following statement in this section: "The authors declare that they have no competing interests."

Author details

[1] Graduate School of Informatics, Kyoto University, Kyoto, Japan.

[2] Center for Innovative Research and Education in Data Science, Kyoto University, Kyoto, Japan.

[3] Independent Researcher

[4] Graduate School of Advanced Science and Engineering, Hiroshima University, Hiroshima, Japan

[5] Academic Center for Computing and Media Studies, Kyoto University, Kyoto, Japan

Received: 10 February 2025 Accepted: 5 November 2025

Published online: 1 January 2027 (Online First: 16 March 2026)

References

Ait Baha, T., El Hajji, M., Es-Saady, Y., & Fadili, H. (2023). The impact of educational chatbot on student learning experience. *Education and Information Technologies*, *29*(8), 10153–10176.

Anderson, L. (2018). Fostering metalinguistic awareness: Role play, pragmatics and L2 literary translation. *Intralinea On Line Translation Journal*, 323–334.

Antón Remírez, S., & Lázaro Ibarrola, A. (2011). Lost in translation? Translation as a valid tool for the EFL classroom. *Huarte de San Juan*. https://academica-e.unavarra.es/handle/2454/9342

Bibauw, S., Van den Noortgate, W., François, T., & Desmet, P. (2022). Dialogue systems for language learning: A meta-analysis. *Language Learning & Technology*, *26*(1). https://lirias.kuleuven.be/3246192?limo=0

Bowker, L., McBride, C., & Marshman, E. (2008). Getting more than you paid for? Considerations in integrating free and low-cost technologies into translator training programs. *Revista Electrónica de Didáctica de La Traducción y La Interpretación*, *1*, 26–47.

Brinegar, M. (2023). Chatbots as a Supplementary Language Learning Tool: Advantages, Concerns, and Implementation. *International Journal of Education and Social Science Research*, *6*(6), 223–230.

Calis, E., & Dikilitas, K. (2012). The Use of Translation in EFL Classes as L2 Learning Practice. *Procedia - Social and Behavioral Sciences*, *46*, 5079–5084.

Casas, J., Tricot, M.-O., Abou Khaled, O., Mugellini, E., & Cudré-Mauroux, P. (2021). Trends & Methods in Chatbot Evaluation. *Companion Publication of the 2020 International Conference on Multimodal Interaction*, 280–286.

Chang, C.-K., & Hsu, C.-K. (2011). A mobile-assisted synchronously collaborative translation–annotation system for English as a foreign language (EFL) reading comprehension. *Computer Assisted Language Learning*, *24*(2), 155–180.

Chuah, K., & Kabilan, M. (2021). Teachers' views on the use of chatbots to support English language teaching in a mobile environment. *International Journal of Emerging Technologies in Learning (iJET)*, *16*(20), 223–257.

Cunningham-Nelson, S., Boles, W., Trouton, L., & Margerison, E. (2019). *A review of chatbots in education: Practical steps forward* (pp. 299–306). AAEE - Annual Conference of Australasian Association for Engineering Education, Brisbane, Australia, AUS. Engineers Australia.

Dagilienė, I. (2012). Translation as a Learning Method in English Language Teaching. *Studies about Languages* , *21*, 124–129.

Deng, X., & Yu, Z. (2023). A Meta-Analysis and Systematic Review of the Effect of Chatbot Technology Use in Sustainable Education. *Sustainability: Science Practice and Policy*, *15*(4), 2940.

Duolingo Team. (2023, March 14). *Introducing Duolingo Max, a learning experience powered by GPT-4*. https://blog.duolingo.com/duolingo-max/

Folse, K. S. (2004). Myths about teaching and learning second language vocabulary: What recent research says. *TESL Reporter*, *37*, 1–13.

Goodman, R. S., Patrinely, J. R., Stone, C. A., Jr, Zimmerman, E., Donald, R. R., Chang, S. S., Berkowitz, S. T., Finn, A. P., Jahangir, E., Scoville, E. A., Reese, T. S., Friedman, D. L., Bastarache, J. A., van der Heijden, Y. F., Wright, J. J., Ye, F., Carter, N., Alexander, M. R., Choe, J. H., … Johnson, D. B. (2023). Accuracy and reliability of chatbot responses to physician questions. *JAMA Network Open*, *6*(10), e2336483.

Granić, A., & Marangunić, N. (2019). Technology acceptance model in educational context: A systematic literature review. *British Journal of Educational Technology: Journal of the Council for Educational Technology*, *50*(5), 2572–2593.

Huang, W., Hew, K. F., & Fryer, L. K. (2022). Chatbots for language learning—Are they really useful? A systematic review of chatbot-supported language learning. *Journal of Computer Assisted Learning*, *38*(1), 237–257.

Huh, M.-H. (2001). Translation strategy in EFL writing. *English Teaching*, *56*, 75–92.

Inderawati, R., Hayati, R., Marlina, R., Novarita, N., Awalludin, A., & Anam, S. (2023). Argumentative Essay and Vocabulary Enrichment of English Students by Utilizing Google Translate. *English Community Journal*, *6*(2), 131–141.

JASP Team. (2024). *JASP (Version 0.18.3)[Computer software]*. https://jasp-stats.org/

Jeon, J. (2023). Chatbot-assisted dynamic assessment (CA-DA) for L2 vocabulary learning and diagnosis. *Computer Assisted Language Learning*, *36*(7), 1338–1364.

Kang, S., & Sung, M.-C. (2024). EFL students' self-directed learning of conversation skills with AI chatbots. *Language Learning & Technology*, *28*(1), 1–19.

Karjo, C. H., & Andreani, W. (2018). Learning Foreign Languages With Duolingo and Memrise. *Proceedings of the 2018 International Conference on Distance Education and Learning*, 109–112.

Kessler, M., Loewen, S., & Gönülal, T. (2023). Mobile-assisted language learning with Babbel and Duolingo: Comparing L2 learning gains and user experience. *Computer Assisted Language Learning*, 1–25.

Kim, H.-S., Cha, Y., & Kim, N. Y. (2021). Effects of AI chatbots on EFL students' communication skills. *Korean Journal of English Language and Linguistics*, *21*, 712–734.

Kim, N.-Y. (2019). A study on the use of artificial intelligence chatbots for improving English grammar skills. *Journal of Digital Convergence*, *17*, 37–46.

Law, L. (2024). Application of generative artificial intelligence (GenAI) in language teaching and learning: A scoping literature review. *Computers and Education Open*, *6*(100174), 100174.

Levy, M., & Stockwell, G. (2013). *CALL dimensions: Options and issues in computer-assisted language learning*. Routledge.

Lin, H. (2015). A meta-synthesis of empirical research on the effectiveness of computer-mediated communication (CMC) in SLA. *Language Learning & Technology*, *19*(2), 85–117.

Liu, X., Zhang, J., Barany, A., Pankiewicz, M., & Baker, R. S. (2024). *Assessing the Potential and Limits of Large Language Models in Qualitative Coding*. 89–103.

Loewen, S., Crowther, D., Isbell, D. R., Kim, K. M., Maloney, J., Miller, Z. F., & Rawal, H. (2019). Mobile-assisted language learning: A Duolingo case study. *ReCALL*, *31*(3), 293–311.

Murtisari, E. T., Bonar, G. J., & Widiningrum, R. (2020). Learning grammar through learning to translate: A means and an end. *The Journal of Asia TEFL*, *17*(2), 715–723.

Nation, P. (1996). The four strands of a language course. *TESOL in Context*, *6*, 7.

Nicholson, R., Bartindale, T., Kharrufa, A., Kirk, D., & Walker-Gleaves, C. (2022). Participatory design goes to school: Co-teaching as a form of co-design for educational technology. *CHI Conference on Human Factors in Computing Systems*. CHI '22: CHI Conference on Human Factors in Computing Systems, New Orleans LA USA. 1–17.

Ouyang, Z., Jiang, Y., & Liu, H. (2024). The effects of Duolingo, an AI-integrated technology, on EFL learners' willingness to communicate and engagement in online classes. *The International Review of Research in Open and Distributed Learning*, *25*(3), 97–115.

Phanwiriyarat, K., Anggoro, K. J., & Chaowanakritsanakul, T. (2025). Exploring AI-powered gamified flipped classroom in an English-speaking course: A case of Duolingo. *Cogent Education*, *12*(1), 1–19.

Rebolledo Font de la Vall, R., & Araya, F. G. (2023). Exploring the benefits and challenges of AI-language learning tools. *The International Journal of Social Sciences and Humanities Invention*, *10*(1), 7569–7576.

Saleh, N. S., Murtaza, S. F., & Baki, N. U. (2021). A Comparative Analysis on Cohesive Markers in Essay Composition of First Language and Second Language: Direct Writing Versus Translation. *International Journal of Linguistics Studies* , *1*(2), 1–10.

Smith, R. (2005). Global English: Gift or curse? *English Today*, *21*(2), 56–62.

Tavakoli, M., Ghadiri, M., & Zabihi, R. (2014). Direct versus translated writing: The effect of translation on learners' second language writing ability. *GEMA Online Journal of Language Studies*, *14*(02), 61–74.

Tekwa, K. (2024). *Machine Translation and Foreign Language Learning*. Springer Nature.

Watkins, R. (2023). Guidance for researchers and peer-reviewers on the ethical use of Large Language Models (LLMs) in scientific research workflows. *AI and Ethics*. *4*(4), 969–974.

Woollaston, S., Flanagan, B., Ocheja, P., Dai, Y., & Ogata, H. (2024). TAMMY: Supporting EFL Translation Practice With an LLM-Powered Chatbot. *Proceedings of the 32nd International Conference on Computers in Education*, *1*, 595–604.

Woollaston, S., Flanagan, B., Ocheja, P., Toyokawa, Y., & Ogata, H. (2025). ARCHIE: Exploring language learner behaviors in LLM chatbot-supported active reading log data with epistemic network analysis. *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, 642–654.

Woollaston, S., Flanagan, B., & Ogata, H. (2024). Chatbots and EFL learning: A systematic review. *Joint Proceedings of LAK 2024 Workshops*, 89–98.

Wu, X., & Gao, Y. (2011). Applying the extended technology acceptance model to the use of clickers in student learning: Some evidence from macroeconomics classes. *American Journal of Business Education*, *4*, 43–50.

Xu, X., Sun, F., & Hu, W. (2025). Integrating human expertise with GenAI: Insights into a collaborative feedback approach in translation education. *System*, *129*(103600), 103600.

Yang, J. (2022). Perceptions of preservice teachers on AI chatbots in English education. *International Journal of Internet, Broadcasting and Communication*, *14*(1), 44–52.

Zambrano, A. F., Liu, X., Barany, A., Baker, R. S., Kim, J., & Nasiar, N. (2023). From nCoder to ChatGPT: From automated coding to refining human coding. In *Communications in Computer and Information Science* (pp. 470–485). Springer Nature Switzerland.

Zhai, C. (2023). A systematic review on artificial intelligence dialogue systems for enhancing English as foreign language students' interactional competence in the university. *Computers and Education: Artificial Intelligence*, *4*, 100134.

Zhang, R., & Zou, D. (2022). Types, purposes, and effectiveness of state-of-the-art technologies for second and foreign language learning. *Computer Assisted Language Learning*, *35*(4), 696–742.

Publisher's Note

The Asia-Pacific Society for Computers in Education (APSCE) remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Research and Practice in Technology Enhanced Learning (RPTEL) **is an open-access journal and free of publication fee.**