frontiers | Frontiers in Education

TYPE Original Research
PUBLISHED 06 February 2025
DOI 10.3389/feduc.2025.1499497

OPEN ACCESS

EDITED BY
Jesus Martin-Vaquero,
University of Salamanca, Spain

REVIEWED BY
Dennis Arias-Chávez,
Continental University, Peru
Ela Luria,
Levinsky College of Education, Israel
Can Mese,
Kahramanmaraş Istiklal University, Türkiye

*CORRESPONDENCE
Audrey K. Kittredge
audrey@duolingo.com

RECEIVED 20 September 2024
ACCEPTED 21 January 2025
PUBLISHED 06 February 2025

CITATION
Kittredge AK, Hopman EWM, Reuveni B, Dionne D, Freeman C and Jiang X (2025) Mobile language app learners' self-efficacy increases after using generative AI. *Front. Educ.* 10:1499497. doi: 10.3389/feduc.2025.1499497

Mobile language app learners' self-efficacy increases after using generative AI

Audrey K. Kittredge*, Elise W. M. Hopman, Ben Reuveni, Danielle Dionne, Cassie Freeman and Xiangying Jiang

Duolingo, Pittsburgh, PA, United States

Introduction: Although generative artificial intelligence (AI) is ubiquitous, there is little research on how it supports self-efficacy (learners' belief that they can perform at a particular level on a specific task). The purpose of these studies was to investigate self-efficacy development in a generative AI-based language learning experience.

Methods: In two studies, learners (N = 385) of French/Spanish used AI-based features offering conversation practice and on-demand explanations in a mobile app (Duolingo) for 1 month. Before and after using the features, learners reported their self-efficacy and other perceptions.

Results: In Study 1, learners who had already used the features felt significantly more prepared to use French/Spanish in real-life situations after 1 month, as did learners in Study 2 who used the features for the first time. Learners in Study 2 also felt significantly more prepared to share their opinions and navigate a city, and reported significantly higher self-efficacy for speaking and understanding grammar and mistakes. Across studies, the majority of learners agreed that the AI-based features effectively supported learning, and reported using their learning outside the app.

Discussion: These results provide the first evidence of enhanced language learning self-efficacy after use of generative AI, building on findings from classroom interventions.

KEYWORDS

language learning, Mobile-assisted language learning (MALL), language self-efficacy, generative AI, learning transfer

1 Introduction

Learning a language is challenging, and supporting learners' motivation during the language learning process is critical (Albalawi and Al-Hoorie, 2021). Motivation for learning results from how much learners *value* learning, as well as their *expectancy* that learning will lead to specific outcomes (Vu et al., 2022). One type of expectancy is *self-efficacy*, learners' belief in the ability to perform at a particular level on a specific task (Bandura, 1994). Self-efficacy plays an important role in theories of learning, and studies across a variety of educational domains, including language learning, reveal significant correlations between self-efficacy, performance, and behaviors that support learning (Zimmerman, 2000; Goetze and Driver, 2022).

Classroom interventions have successfully increased students' language learning self-efficacy with communication-focused tasks, constructive feedback that highlights learners' success, and explicit teaching of language learning strategies (Raoofi et al., 2012; Graham, 2022). Although generative AI models such as ChatGPT are well-positioned to provide

real-time communication practice that could increase self-efficacy, there is little research on this topic (Han, 2024; Law, 2024). This study aims to help fill this gap in the research literature.

2 Self-efficacy in language learning

Self-efficacy plays an important role in learning theory (Zimmerman, 2000; Wigfield and Eccles, 2000). Individuals form self-efficacy beliefs based on their past performance, and these self-efficacy beliefs exert a strong influence on behaviors such as learning goals and strategies, which impact subsequent performance (Talsma et al., 2018). Performance is thought to influence self-efficacy, and self-efficacy is thought to influence performance (Vu et al., 2022), in a mutually reinforcing cycle (see Figure 1).

Consistent with these theories, empirical evidence from a variety of countries in Africa, Asia, Europe, the Middle East, and North America suggests that self-efficacy plays an important role in second language learning. Self-efficacy correlates with actual task performance in a variety of domains (Honicke and Broadbent, 2016), including second language learning (Yang and Lian, 2023; Young Kyo, 2022), with a larger effect size than for other motivational variables (Schneider and Preckel, 2017; Goetze and Driver, 2022). Longitudinal data show that when learners' self-efficacy increases, performance tends to increase as well (Bernacki et al., 2015; Vu et al., 2022). Self-efficacy is also correlated with goal-directed behaviors that promote learning (Zimmerman, 2000). For example, high self-efficacy learners tend to use more language learning strategies (Raoofi et al., 2012; Wang and Sun, 2020).

What cultivates language learning self-efficacy? Although most research on this topic is correlational, a number of classroom interventions in Asia, Europe, the Middle East, and North America have increased students' self-efficacy. These interventions feature communication-focused tasks such as writing and speaking activities (Abdelhalim, 2024; Leeming, 2017; Mills, 2009; Goetze and Driver, 2022), constructive feedback that highlights students' success (Xu et al., 2022; Li et al., 2023), and explicit teaching of learning strategies (Chen, 2022; Milliner and Dimoski, 2024). Taken together, these studies suggest that new interventions with these same elements could support the development of language learning self-efficacy.

3 Generative AI to support self-efficacy

Given the prominence of communication-focused tasks in interventions that support self-efficacy, it is possible that real-time conversation practice provided by generative AI could enhance self-efficacy as well. Research conducted prior to the release of Open AI's GPT in Asia, Europe, the Middle East and North America found that chatbots simulating human interaction can enhance motivational constructs related to self-efficacy, such as speaking confidence and willingness to communicate in the language (Du and Daniel, 2024; Xiao et al., 2023). However, experimental evidence on ChatGPT and other recent generative AI is still limited (Han, 2024; Law, 2024).

One practically relevant context for investigating the impact of learning experiences supported by generative AI is widely used language learning apps (Tommerdahl et al., 2024), which have recently incorporated this new technology (Godwin-Jones, 2024). Although rigorous research on language learning apps is limited, several apps have demonstrated efficacy relative to a control group (Tommerdahl et al., 2024). In particular, three apps have been investigated in multiple experimental studies: Rosetta Stone, Memrise, and Duolingo.

When the application Rosetta Stone is combined with classroom instruction, studies in China (Bai, 2024; Fan, 2023) and the US (Harper et al., 2021) find that this leads to better learning outcomes than classroom instruction alone. Memrise, a vocabulary learning app, leads to comparable or better learning outcomes than classroom instruction in China (Wang et al., 2023), Iran (Shamshiri et al., 2023), and Vietnam (Nguyen et al., 2023). Duolingo, a language learning app studied in over 300 articles (Shortt et al., 2023), produces learning outcomes that are comparable to or better than classroom instruction in China (Qiao and Zhao, 2023), Colombia (García Botero et al., 2021), Russia (Pichugin et al., 2023), and the US (Rachels and Rockinson-Szapkiw, 2018). Duolingo also yields learning outcomes that are comparable to or better than other language learning apps, in Russia (Pichugin et al., 2023) and the US (Kessler et al., 2023), and is effective when used by learners outside of formal education settings in various countries (Jiang et al., 2021; Jiang et al., 2024b). Duolingo is highly relevant to many learners, as the most downloaded digital

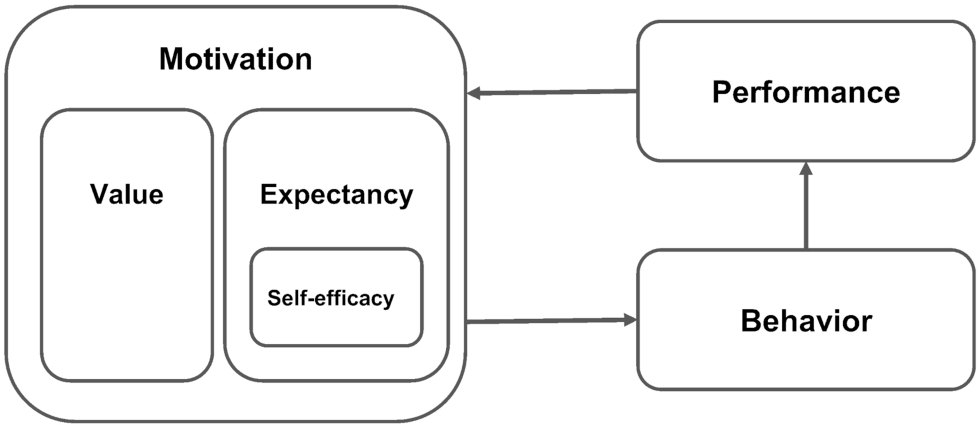

FIGURE 1
Theoretical relationship between motivation (including self-efficacy), behavior, and performance.

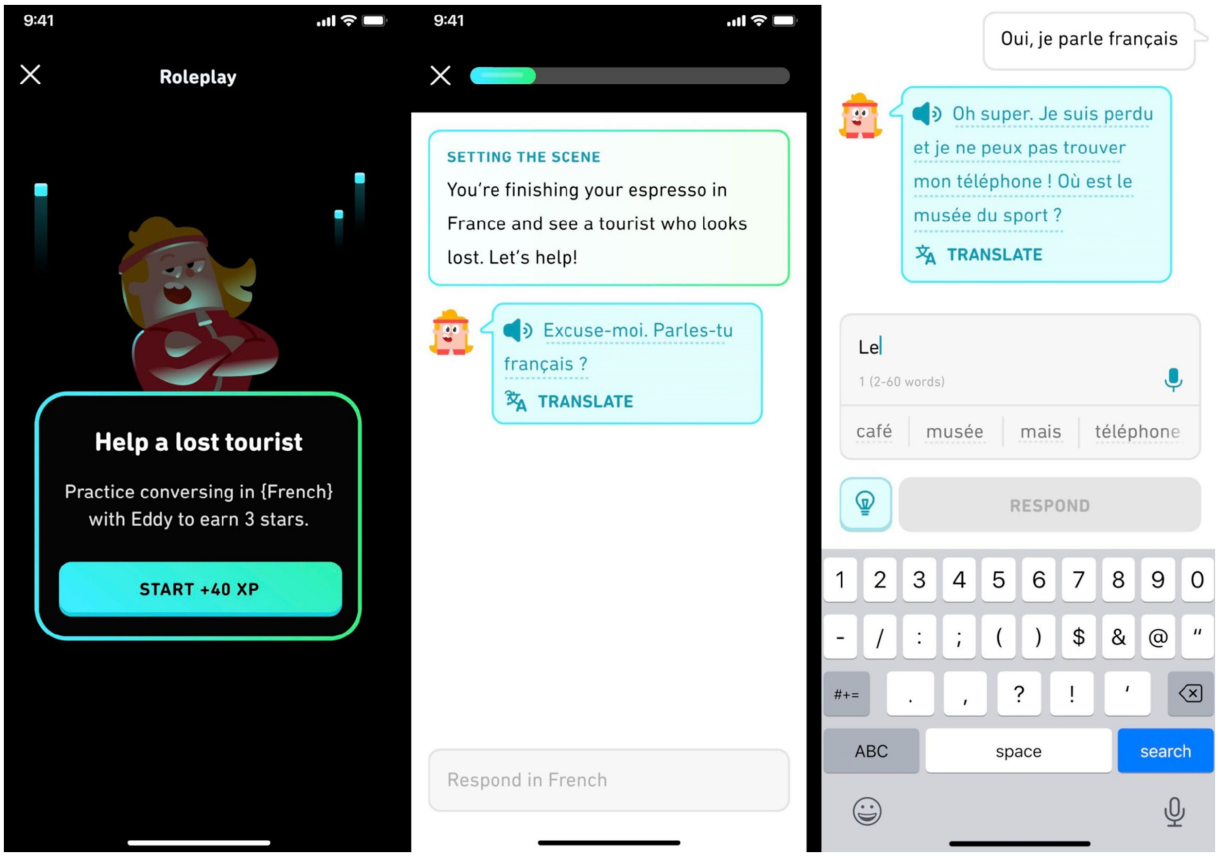

FIGURE 2
Example screens of the Duolingo roleplay feature (Version from August 2023). Translations of the French text are "Excuse me. Do you speak French?" (middle screen), "Yes, I speak French" and "Oh great. I am lost and I cannot find my phone! Where is the sports museum?" (right screen).

language learning product in the world (Statista, 2024). Moreover, motivating learners is a key component of Duolingo's research-based teaching method. Many different elements, such as short lessons, celebrations of learner success, rewards for accuracy and continued app use, and social features work in concert to increase learners' expectancy of positive learning outcomes (Freeman et al., 2023).

The present research focuses on Duolingo, with the following rationale: (1) A greater number of studies have demonstrated Duolingo's learning efficacy, compared to other apps. (2) These studies suggest that Duolingo learning outcomes are comparable to learning outcomes in other apps and classrooms, (3) in settings both within and outside of formal schooling. (4) The app's widespread use increases the relevance of any research conducted on it, and (5) Duolingo's teaching method grounded in research on learning and motivation is well suited to supporting the development of self-efficacy.

Several Duolingo features that use Open AI's GPT-4 (OpenAI, 2023b) were released as part of a higher subscription tier[1] (Duolingo Team, 2023). In the "Roleplay" feature (Figure 2), learners practice written or spoken real-world conversation with characters whose live responses are generated by GPT-4, in scenarios like taking a taxi to the airport. Learners are offered translation hints, helpful phrase suggestions, feedback on accuracy, and tips for future conversations, all of which are generated live by GPT-4. In Duolingo's "Explain My Answer" feature (Figure 3), learners can optionally tap on a button after certain exercises to enter a chat, where GPT-4 is used to generate an explicit explanation of why the learners' answer was right or wrong, along with examples or further clarification.

Duolingo's AI-enabled learning experiences contain the same core elements that have been shown to enhance self-efficacy in the classroom: communication-focused tasks, constructive feedback that highlights learners' success, and explicit teaching of language learning strategies. These similarities, in addition to the focus on motivation in Duolingo's teaching method (Freeman et al., 2023), suggest that the experiences provided by the AI-based features may enhance learners' self-efficacy. However, there is no experimental evidence to date that demonstrates an increase in self-efficacy after using Duolingo, or any other language learning app that incorporates generative AI.

4 Research questions and overview of the studies

The goal of the current research is to explore the development of language learning self-efficacy in learning experiences supported by

1 This higher tier included features that are non-essential for learning (unlimited hearts, no ads, personalized review), as well as GPT-4 features that were too expensive at the time to be included in the free version of Duolingo. The subscription was called "Duolingo Max" at the time of the study.

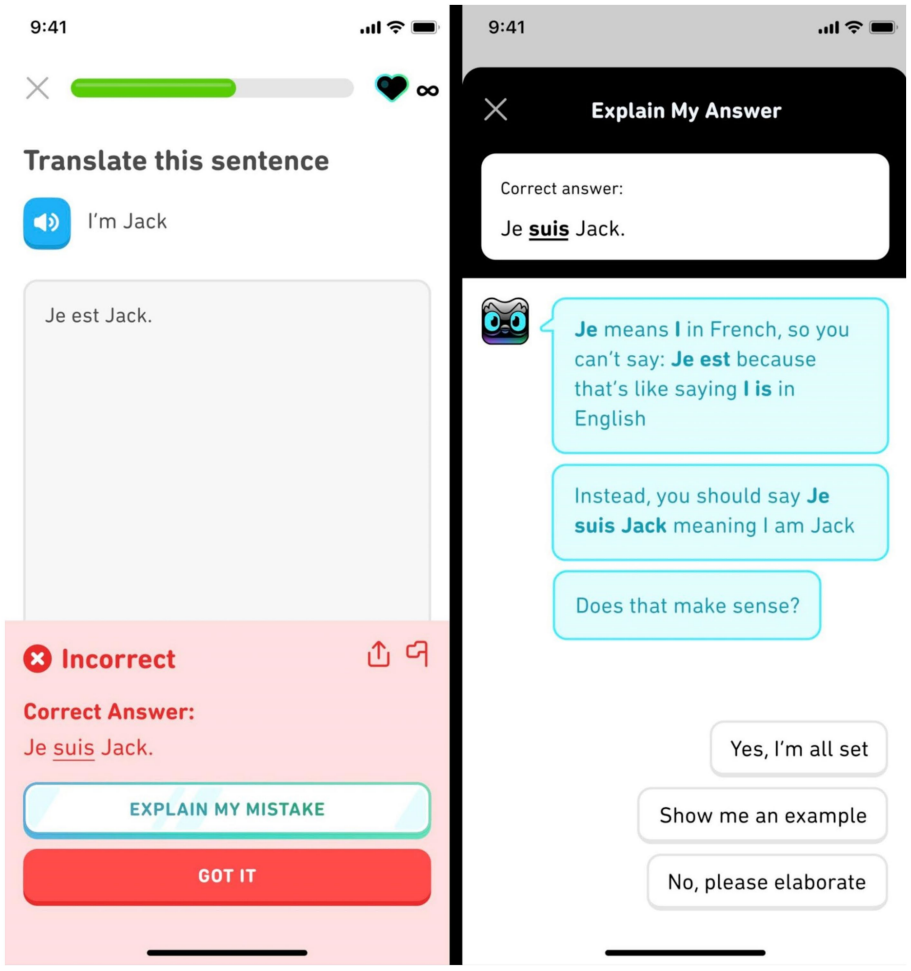

FIGURE 3
Example screens of the Duolingo explain my answer feature (Version from August 2023).

generative AI. A total of 385 learners in the Duolingo French and Spanish courses used AI-based features for 1 month. These learners took a survey before and after using the features, to answer the following exploratory Research Questions:

> *RQ1*. Are there significant increases in learners' self-efficacy for language skills after 1 month of using AI-based features?
>
> *RQ2*. What percentage of learners believe that AI-based features effectively support language learning?
>
> *RQ3*. Do learners use what they learned from AI-based features outside the app, and if so, how?

The present research also investigated AI-based feature novelty. Economic and computational models of human decision-making hypothesize that novelty has a high value, because gathering new information could benefit future decisions (Wittmann et al., 2008). Consistent with these theories, research in psychology demonstrates that novelty leads to enhanced attention (Kagan, 2009) as well as better learning and memory (Barto et al., 2013), and research in marketing suggests that customers' perceived value of novel products is higher (Blut et al., 2023; Sánchez-Fernández and Iniesta-Bonillo, 2007; Leroi-Werelds, 2019). Novelty effects are prevalent in perceptions of technology: Technologies with higher perceived novelty are associated with more positive attitudes and higher intentions of use (Wells et al., 2010). In interactions with pre-GPT chatbots, learners report higher interest in language learning when the chatbot was novel (Fryer et al., 2017; Fryer et al., 2019). Research in business and medicine also suggests that reduced cost of products and medicines can lead to more positive perceptions (Blut et al., 2023; Lee et al., 2020; Wang et al., 2013). On the other hand, research in economics demonstrates that people value things that they own more (the "endowment effect"; Marzilli Ericson and Fuster, 2014), suggesting that cost associated with ownership could also lead to positive perceptions.

To determine whether AI-based feature novelty and cost could impact learners' self-efficacy, Study 1 was conducted with pre-existing subscribers who had already been using the AI-based features, while Study 2 was conducted with learners who had been using the free version of Duolingo and were granted complimentary access to the features.

5 Study 1

5.1 Methodology

5.1.1 Participants

The participants were 280 pre-existing subscribers who met the following inclusion criteria:

1 They consented to be contacted by Duolingo for research purposes.
2 They were enrolled in the French or Spanish course for English speakers (the subscription was only offered in those courses at the time of the study).
3 They completed no more than the first or second section of the "basic" (*Common European Framework of Reference* A1 level; Council of Europe, 2001) course content, in line with the majority of subscribers at the time of the study.
4 Their average daily app use was between 15 minutes (the time to complete several Duolingo lessons) and 27 minutes (the 75th percentile of subscribers' daily app usage).
5 They were using Duolingo on iOS (the subscription was only available on this platform at the time of the study).
6 They were located in one of six English-speaking countries (the subscription was only available in those countries at the time of the study).
7 They self-reported their age as 18 or older.

5.1.2 Research tools

5.1.2.1 Background questionnaire

The background questionnaire asked participants' age, as well as their reasons for learning the language, knowledge of the language prior to using Duolingo, and use of other programs/apps besides Duolingo to learn the language. The full questionnaire and questionnaire response data can be found in Appendices A, B, respectively.

5.1.2.2 Pre-survey: self-efficacy

The pre-survey (for full text, see Appendix A) collected data on participants' self-efficacy in the language they were learning. Rather than drawing on another published scale, 17 different items were developed to assess self-efficacy on a wide range of different language skills and language use scenarios that could be relevant to learning on Duolingo. In line with previous research on self-efficacy, these were statements of confidence in one's own ability to achieve a performance goal (Talsma et al., 2018). Each statement was rated on a 6-point bipolar Likert-type scale.

While data were collected on all 17 items, subsequent user experience research found that the following seven statements were most relevant to Duolingo learners' use of the Roleplay and Explain My Answer features. To increase statistical power, only responses to these items were analyzed to answer RQ1:

- I feel prepared to use French/Spanish in real-life situations.
- I feel prepared to use French/Spanish to ask for directions and navigate a new city.
- I feel prepared to use French/Spanish to share my opinions with others.
- I'm confident in my ability to understand French/Spanish grammar.
- I'm confident in my ability to understand the mistakes I make in French/Spanish.
- I'm confident in my ability to understand spoken French/Spanish.
- I'm confident in my ability to speak in French/Spanish.

5.1.2.3 Post-survey: self-efficacy and perceptions of the AI-based features

The post-survey (for full text, see Appendix A) included the same 17 self-efficacy statements as on the pre-survey, as well as 18 additional items developed to assess participants' perceptions of the AI-based features.

Sixteen of the 18 items referred to a wide range of different language skills and language use scenarios relevant to learning on Duolingo. While data were collected on all these items, the analysis to answer RQ2 focused on the following eight statements for the same reasons as mentioned above (maximizing relevance and statistical power):

- Duolingo Max prepares me for real-world situations.
- Duolingo Max helps me learn from my mistakes.
- Duolingo Max helps me understand conversations.
- Duolingo Max helps me understand grammar.
- Duolingo Max helps me learn to speak.
- "Roleplay" prepares me to use French/Spanish in real-world situations.
- "Explain My Answer" helps me better understand grammar.
- "Explain My Answer" helps me learn from my mistakes.

Duolingo Max was the name of the subscription at the time of the study.

One of the 18 items asked about learners' experiences using their learning outside the app, and another elicited general feedback on the study experience. We only analyzed answers to the former, which was relevant to RQ3:

- Have you used what you learned with Duolingo Max outside of the app? (Yes/No).
- [*Only displayed if 'Yes' was selected*] Please describe the situation (optional).

5.1.3 Data collection procedure

Data were collected between August 30, 2023 and November 17, 2023 on a rolling basis. Figure 4 shows the steps of the study and the number of participants involved, described in more detail below.

Based on a power analysis with a small estimated effect size (0.2), the aim was for 150 participants to complete the full, month-long study. To reach our target sample size, emails soliciting participation were sent in multiple recruitment waves to 1,421 subscribers who met inclusion criteria 1–6 (see Participants section). As the subscriber population was skewed towards Spanish learners at the time of the study, 1,030 of invited subscribers were in the Spanish course for English speakers, and 391 were in the French course for English speakers.

Participants were asked to complete the background questionnaire. Those who were 18 years or older were invited to take the pre-survey, and

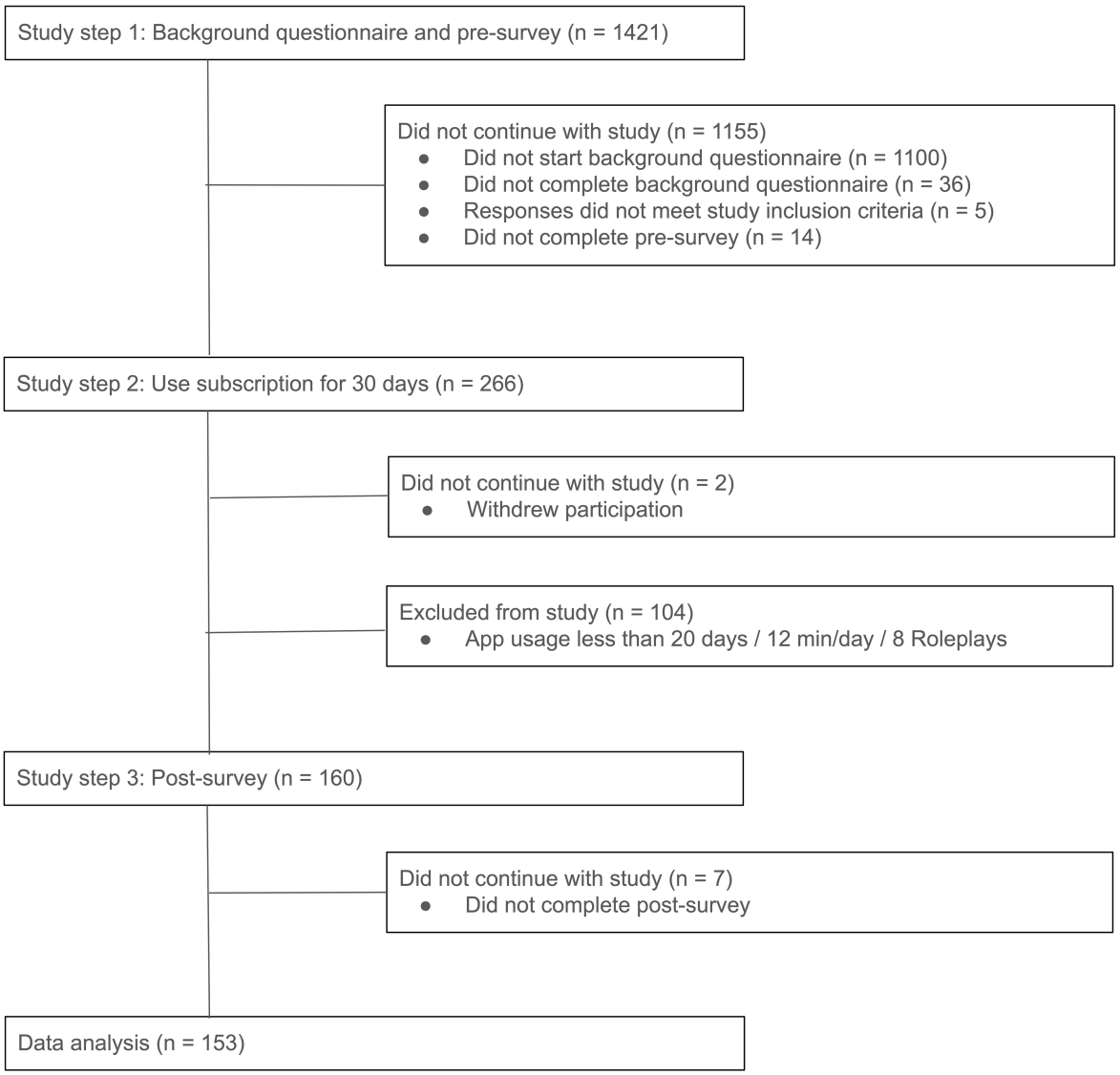

FIGURE 4
Study 1 Steps with numbers of participants at each step.

then use Duolingo for 30 days. During the 30 days they were asked to use Duolingo for at least 15 minutes per day, including AI-based features such as Roleplay and Explain My Answer, and to complete at least 15 Roleplays by the end of the 30 days. Participants received weekly reminder emails summarizing their app usage, and were excluded when there was no chance of them meeting minimal usage requirements (using the app for at least 20 days by the end of the study, average app usage excluding Roleplays above 12 minutes per day, more than 8 Roleplays in total).

After the 30 days, participants were invited to take the post-survey if they had been active for at least 25 days, completed at least 15 Roleplays, and had at least 12 minutes of app usage per day on average. If they had less usage (active for 20–24 days, completed 8–14 Roleplays, at least 12 minutes usage per day), they were given an additional 5 days to use the app and then invited to the post-survey. Participants who had even less app usage (active for 20–24 days, completed 8–14 Roleplays, less than 12 min usage per day) were partially compensated $15 and not invited to take the post-survey.

After completing the post-survey, participants received $40 as promised in the initial invitation email, as well as a surprise complimentary 31-day subscription extension.

Treatment attrition (participants not meeting app usage requirements) was 40%. The background questionnaire responses of excluded participants and those who completed the study were similar (see Tables B1, B2), and these groups did not differ significantly in pre-survey self-efficacy responses ($p > 0.05$ in Mann–Whitney U tests; see Tables C1, C2).

5.1.4 Data analysis

To answer RQ1, participants' responses to the self-efficacy items were analyzed to see if there was a change from the pre-survey to the post-survey. Likert scale ratings were converted to a binary "disagree" or "agree" rating ("strongly disagree," "disagree," or "slightly

disagree" = 0, "slightly agree," "agree," or "strongly agree" = 1). Although we acknowledge shortcomings of binarizing Likert scale data (e.g., data loss and possible error rate inflation, Royston et al., 2006), there is precedent for this approach when the response can be considered binary (Harpe, 2015). A logistic regression analysis was conducted to determine if participants were more likely to agree with a self-efficacy statement on the post-survey than the pre-survey, given their pre-survey responses and app usage:

> Post-survey response ~ Intercept + Pre-survey response + number of Roleplays completed + number of Explain My Answers shown.

We mean-centered app usage predictors, and checked that multicollinearity, sample size requirements, and linearity of the logit assumptions held. Since we conducted a separate logistic regression for each of the 7 self-efficacy statements, we corrected for multiple comparisons (Benjamini and Hochberg, 1995).

To answer RQ2, participants' responses to the feature perception items were analyzed. As in the RQ1 analysis, the Likert scale ratings were converted to a binary "disagree" or "agree" rating, and the percentages of participants who agreed with each statement were calculated. A one-sided binomial test was used to check if each percentage was significantly greater than 80%.

To answer RQ3, the percentage of "yes" responses to the question about using learning outside the app was calculated, and themes were identified in learners' descriptions of using learning outside the app. Following recent use of large language models for thematic analysis (Morgan, 2023; Moorhouse and Kohnke, 2024), Glean's "AI Summarization" feature (Glean, 2024), which used GPT-4 (OpenAI, 2023b), GPT-3.5-Turbo (OpenAI, 2023a), and text-embedding-ada-002 (OpenAI, 2022), was given the following two prompts to identify themes appearing in 10% of participants' responses: (1) "I need to know the most common situations in which Duolingo learners used what they learned outside the app. Here is a document with learner descriptions of these situations; each bullet point represents one learner's description. Can you write a summary of the main themes in these learner descriptions, making sure that each theme appears in the descriptions of at least 10 learners?" (2) "Thank you! Can you please provide evidence for each theme summarized above, by citing 3 learners' descriptions that contain the theme?"

Glean produced a set of themes with examples (see Tables C6, E6). To reduce inaccuracy and insufficient nuance (Morgan, 2023), a research assistant checked that all Glean-generated themes could be matched with at least 10 learner descriptions, and that Glean-generated text did not contain any hallucinations. Any inaccuracies were removed.

5.2 Results

5.2.1 Learners' self-efficacy increased significantly

To answer RQ1, we first calculated descriptive statistics of self-efficacy ratings. Of the 15–45% of learners who disagreed with the statements on the pre-survey, 39–69% changed their rating to "agree" on the post-survey (see Figure 5). Of the 55–85% of learners who agreed with the statements on the pre-survey, 88–97% still agreed on the post-survey (see Figure 6). Taken together, these results suggest that many participants maintained or adopted positive self-efficacy beliefs (see Table C3). Indeed, regression analyses (see Table 1) showed that participants who disagreed with the statement "I feel prepared to use French/Spanish in real-life situations" on the pre-survey were significantly more likely to agree with it on the post-survey (Intercept Coefficient = 0.79, $z = 2.56$, $p < 0.05$). Participants who initially agreed with this statement were still likely to agree with it on the post-survey (Pre-survey Coefficient = 1.44, $z = 3.18$, $p < 0.01$).

However, participants were not significantly more likely to agree with the other six self-efficacy statements on the post-survey (see

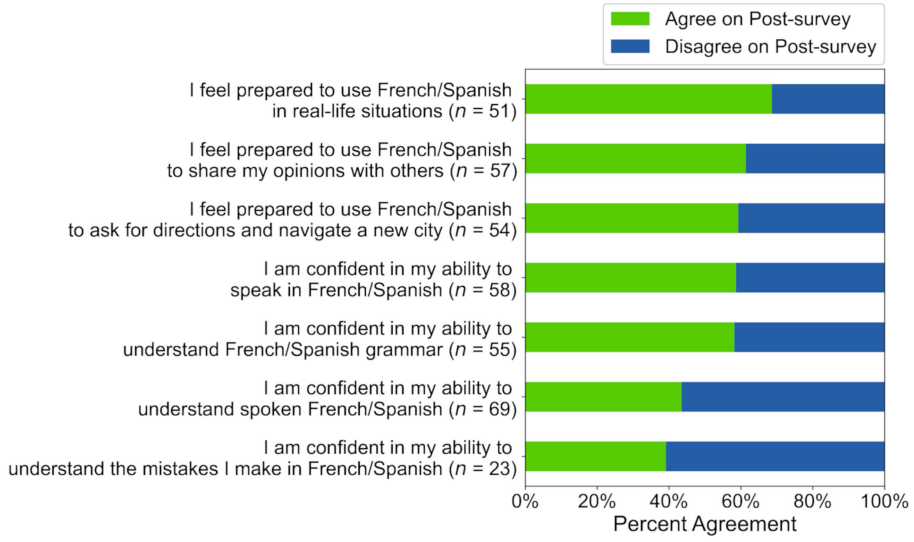

FIGURE 5
Post-survey agreement for participants who disagreed on the pre-survey in Study 1.

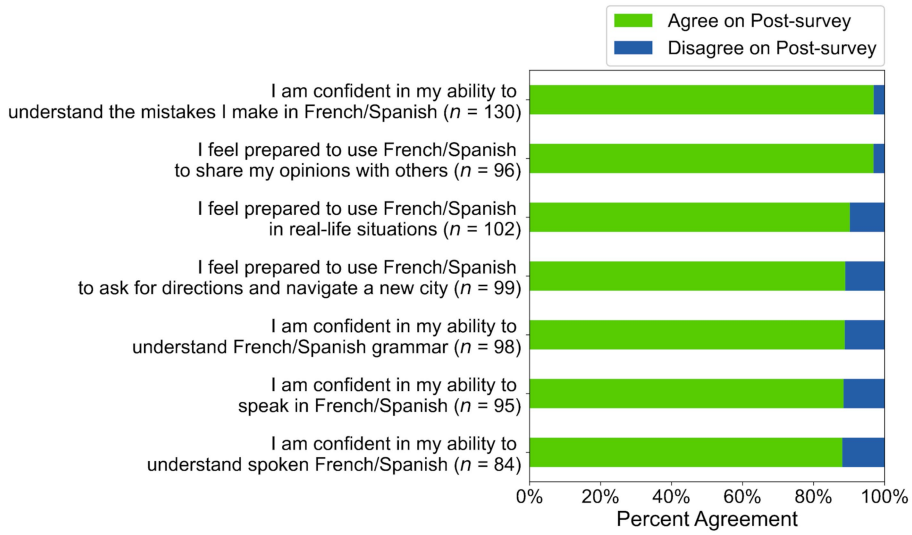

FIGURE 6
Post-survey agreement for participants who agreed on the pre-survey in Study 1.

Table 1). Although participants who initially agreed with these statements were still likely to agree with them on the post-survey (Pre-survey Coefficients all $p < 0.01$ or $p < 0.001$ in these regression models; see Table 1), participants who initially disagreed with them did not demonstrate significant increases in their self-efficacy on the post-survey ($p > 0.05$ for Intercept Coefficients in these regression models; see Table 1).

Participants' average days active ($M = 31.24$, $SD = 2.11$, range = 24–36) and daily average minutes of app usage ($M = 30.92$, $SD = 15.28$, range = 13–99) suggests that they did slightly more than required to stay in the study (see Table B3). However, Roleplay and Explain My Answer feature use did not significantly predict self-efficacy ratings on the post-survey (relevant coefficients in all regressions were not significantly different from 0; see Table 1).

5.2.2 Learners believed AI-based features were effective

To answer RQ2, we calculated the percentage of participants who agreed with statements about AI-based feature efficacy. Agreement rates were significantly above 80% for all eight statements (see Table C5) and 7 of 8 statements had agreement rates numerically above 90% (see Figure 7), suggesting that participants thought the features helped them meet a variety of learning goals.

5.2.3 Learners used what they learned outside the app

To answer RQ3, we analyzed the percentage of learners who said they used what they learned outside the app, as well as themes across these situations. Of the 153 participants, 111 (73%) said they used what they learned with the AI-based features outside the app, and 95 (86%) of the 111 provided descriptions of real-life situations in which they used the language. Table 2 lists the human-verified themes in these participants' responses (see Table C6 for Glean-generated text). Participants described a large variety of real-world contexts, suggesting that many participants used their language skills outside the app and believed they were transferring their learning to real life contexts. Moreover, the fact that so many participants practiced their learning outside the app gave them the ability to truly test their language skills, strengthening the validity of their self-reported self-efficacy ratings and perceptions of the AI-based features.

6 Study 2

6.1 Methodology

6.1.1 Participants

The participants were 302 learners who met the same criteria as the learners in Study 1, except that they did not currently have paid access to the AI-based features as part of a subscription, and their location was not restricted to English-speaking countries.

6.1.2 Research tools

Study 2 used the same research tools as in Study 1.

6.1.3 Data collection procedure

1,111 Duolingo learners who met the eligibility criteria were invited to the study: 555 in the French course for English speakers, and 556 in the Spanish course for English speakers. The same data collection window and procedure applied as in Study 1, except that participants were granted complimentary access to the subscription after completing the pre-survey, so they could use it for the study. Figure 8 shows the study steps with numbers of participants.

There was less treatment attrition in Study 2 (19%), compared to 40% in Study 1. Participants who completed the study tended to be older (19% 18–34 years, 81% 35+ years) than those excluded from the study (27% 18–34, 73% 35+), but other background questionnaire responses in these groups were similar (see Tables D1, D2), and their pre-survey self-efficacy statement responses did not differ significantly ($p > 0.05$ in Mann–Whitney U tests; see Tables E1, E2).

TABLE 1 Full regression output for Study 1 (N = 153).

Statement	Regression formula	Term	Coeff	SE	z	p value
"I feel prepared to use French/Spanish in real-life situations"	Post-survey ~ Intercept + Pre-survey + RP + EMA	Intercept	**0.79**	**0.31**	**2.56**	**p < 0.05**
		Pre-survey	**1.44**	**0.45**	**3.18**	**p < 0.01**
		RP	−0.01	0.01	−0.48	0.70
		EMA	0.004	0.01	0.68	0.63
"I feel prepared to use French/Spanish to ask for directions and navigate a new city"	Post-survey ~ Intercept + Pre-survey + RP + EMA+ RP^2	Intercept	0.39	0.29	1.35	0.28
		Pre-survey	**1.79**	**0.44**	**4.04**	**p < 0.001**
		RP	0.10	0.06	1.83	0.17
		EMA	−0.003	0.01	−0.54	0.70
		RP^2	−0.001	0.001	−1.92	0.17
"I feel prepared to use French/Spanish to share my opinions with others"	Post-survey ~ Intercept + Pre-survey + RP + EMA	Intercept	0.49	0.28	1.74	0.20
		Pre-survey	**2.98**	**0.65**	**4.59**	**p < 0.001**
		RP	0.003	0.02	0.15	0.93
		EMA	0.01	0.01	0.84	0.54
"I'm confident in my ability to understand spoken French/Spanish"	Post-survey ~ Intercept + Pre-survey + RP + EMA + EMA^2	Intercept	−0.28	0.27	−1.06	0.43
		Pre-survey	**2.31**	**0.44**	**5.29**	**p < 0.001**
		RP	−0.01	0.01	−1.00	0.45
		EMA	−0.03	0.01	−1.88	0.17
		EMA^2	0.0002	0.00	1.60	0.22
"I'm confident in my ability to speak in French/Spanish"	Post-survey ~ Intercept + Pre-survey + RP + EMA + EMA^2	Intercept	0.53	0.33	1.59	0.22
		Pre-survey	**1.65**	**0.43**	**3.86**	**p < 0.01**
		RP	−0.002	0.01	−0.13	0.93
		EMA	−0.03	0.02	−1.34	0.28
		EMA^2	0.0003	0.00	1.39	0.28
"I'm confident in my ability to understand French/Spanish grammar"	Post-survey ~ Intercept + Pre-survey + RP + EMA	Intercept	0.46	0.30	1.54	0.22
		Pre-survey	**1.77**	**0.43**	**4.10**	**p < 0.001**
		RP	0.01	0.01	0.66	0.63
		EMA	0.02	0.01	1.96	0.17
"I'm confident in my ability to understand the mistakes I make in French/Spanish"	Post-survey ~ Intercept + Pre-survey + RP + EMA	Intercept	−0.24	0.47	−0.51	0.70
		Pre-survey	**3.95**	**0.70**	**5.67**	**p < 0.001**
		RP	−0.001	0.02	−0.07	0.95
		EMA	0.02	0.01	1.59	0.22

Coeff, Ceofficient; SE, Coefficient Standard Error; RP, mean-centered number of Roleplays completed; RP^2, RP squared then mean-centered; EMA, mean-centered number of times the Explain My Answer feature was shown; EMA^2, EMA squared then mean-centered. Values for Coefficients with $p < 0.05$ are bolded.

6.1.4 Data analysis

The same analyses were used as in Study 1, except that the first prompt to Glean was slightly different. In the portion of the prompt which specified that each theme should be present in 10% of learner responses ("…making sure that each theme appears in the descriptions of at least 10 learners"), "10 learners" was changed to "12 learners" to account for the larger sample size in Study 2.

6.2 Results

6.2.1 Learners' self-efficacy increased significantly

To answer RQ1, we calculated descriptive statistics and conducted regression analyses on self-efficacy statement ratings. Of the 29–46% of learners who initially disagreed with the statements, 61–79% changed their rating to "agree" on the post-survey (see Figure 9), and of the 54–79% of learners who agreed with the statements on the pre-survey, 93–98% still agreed on the post-survey (see Figure 10), suggesting that most participants maintained or adopted positive self-efficacy beliefs (see Table E3).

Regression analyses confirmed this interpretation. For participants who initially disagreed on the pre-survey, there was a significant increase in self-efficacy for 6 of 7 statements: "I feel prepared to use French/Spanish in real-life situations" (Intercept Coefficient = 1.04, $z = 3.88$, $p < 0.001$), "I feel prepared to use French/Spanish to ask for directions and navigate a new city" (Intercept Coefficient = 1.66, $z = 4.27$, $p < 0.001$), "I feel prepared to use French/Spanish to share my opinions with others" (Intercept Coefficient = 1.33, $z = 5.01$, $p < 0.001$), "I'm confident in my ability to speak in French/Spanish"

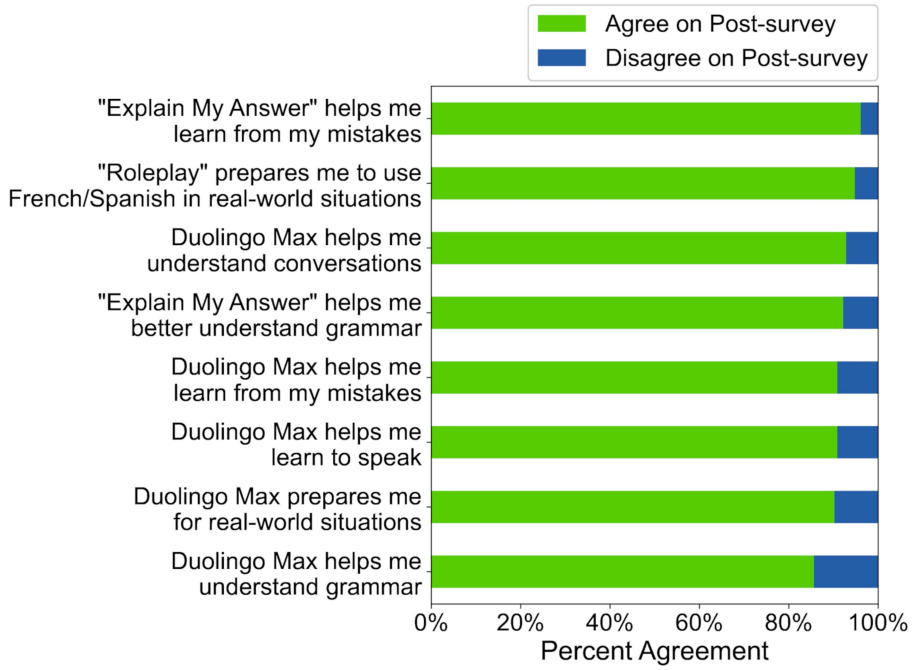

FIGURE 7
Agreement rates for 8 statements on the post-survey in study 1 (N = 153). "Duolingo Max" was the name of the higher subscription tier at the time of the study.

TABLE 2 Summary of AI-generated, human verified themes in Study 1 participants' descriptions of using what they learned outside of Duolingo (N = 95).

Theme	Examples
Travel	Ordering food, asking for directions, making reservations, navigating through different cities.
Conversations with native speakers	With friends, family members, coworkers, or strangers. In various settings such as restaurants, hotels, or on the street.
Workplace communication	With colleagues, patients, and clients who speak the language.
Social interactions	With friends, family, or roommates who are native speakers or also learning the language. Participating in language learning groups or classes.
Service and retail	Ordering at restaurants, speaking with service staff, or helping customers find items in stores.
Remote correspondence	Writing to others, composing emails, texting, or making phone calls.
Education and tutoring	Tutoring students, helping children with homework, or engaging with teachers.

(Intercept Coefficient = 0.94, z = 4.04, $p < 0.001$), "I'm confident in my ability to understand French/Spanish grammar" (Intercept Coefficient = 0.70, z = 2.99, $p < 0.01$), and "I'm confident in my ability to understand the mistakes I make in French/Spanish" (Intercept Coefficient = 0.81, $z = 2.53$, $p < 0.05$). Participants who initially agreed with these same statements were still significantly more likely to agree with them on the post-survey: They felt prepared to use French/Spanish in real-life situations (Pre-survey Coefficient = 1.95, $z = 4.26$, $p < 0.001$), to ask for directions and navigate a new city (Pre-survey Coefficient = 1.74, $z = 3.58$, $p < 0.01$), and to share opinions with others (Pre-survey Coefficient = 1.29, $z = 3.05$, $p < 0.01$). They also felt confident in their ability to speak in French/Spanish (Pre-survey Coefficient = 2.86, $z = 4.58$, $p < 0.001$), understand French/Spanish grammar (Pre-survey Coefficient = 2.19, $z = 5.06$, $p < 0.001$), and understand the mistakes they made in French/Spanish (Pre-survey Coefficient = 3.35, $z = 5.03$, $p < 0.001$).

However, participants who initially disagreed with the statement "I am confident in my ability to understand spoken French/Spanish" did not demonstrate a significant increase in their self-efficacy on the post-survey ($p > 0.05$ for the Intercept Coefficient; see Table 3), although participants who initially agreed with the statement were still likely to agree with it on the post-survey (Pre-survey Coefficient = 2.19, $z = 5.48$, $p < 0.001$). Taken together, these results suggest a significant increase in self-efficacy for a variety of language skills after using the AI-based features.

Like in Study 1, the participants who completed the study met or exceeded the study's app usage requirements (average days active $M = 30.58$, $SD = 1.98$, range = 22–36; daily average minutes of app usage $M = 32.86$, $SD = 15.84$, range = 13–128; see Table D3 for full results). However, Roleplay and Explain My Answer feature use did not significantly predict self-efficacy ratings on the post-survey (coefficients in relevant regressions were not significantly different from 0; see Table 3).

6.2.2 Learners believed AI-based features were effective

To answer RQ2, we calculated the percent of participants who agreed with statements about AI-based feature efficacy. Significantly more than 80% of participants agreed with all eight statements (see

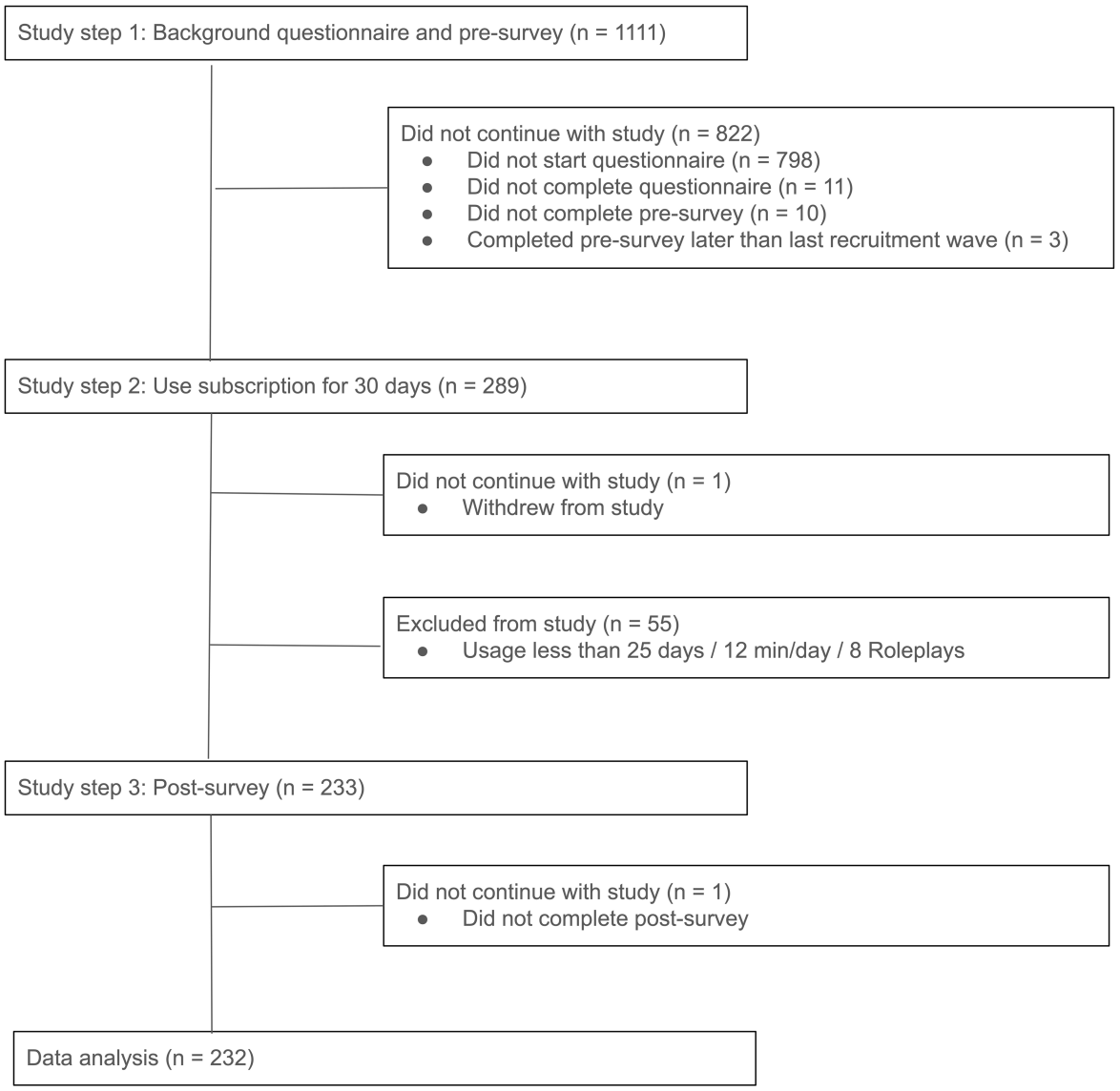

FIGURE 8
Study 2 steps with numbers of participants at each step.

Table E5), and all statement agreement rates were numerically above 90%, suggesting that a large majority of participants thought the AI-based features effectively supported learning (see Figure 11).

6.2.3 Learners used what they learned outside the app

To answer RQ3, we calculated the percentage of learners who said they used what they learned outside the app, as well as themes reported across the use situations. Of the 232 participants, 145 (63%) said they used what they learned with the AI-based features outside the app, and 121 (83%) of the 145 provided descriptions of the real-life situations. These percentages suggest that like in Study 1, participants perceived transfer of learning from the app to use of language skills in real-life.

Table 4 lists themes in participants' responses (see Table E6 for Glean-generated text). The "Travel," "Workplace communication," "Educational support and settings," "Daily errands and services," "Family interaction," and "Casual conversations" themes correspond to the Study 1 themes "Travel," "Workplace communication," "Education and tutoring," "Service and retail," "Social interactions," and "Conversations with native speakers." Other themes showed significant overlap across studies: social media and texting in the theme "Casual conversations" from Study 2 is similar to the "Remote correspondence" theme in Study 1, and the theme "Cultural engagement" in Study 2 was present for some learners in Study 1 who mentioned using media.

7 Discussion

The goal of the present studies was to assess changes in self-efficacy of French and Spanish learners who engaged in learning experiences enabled by generative AI. Study 1 investigated whether there were significant increases in pre-existing, paying subscribers'

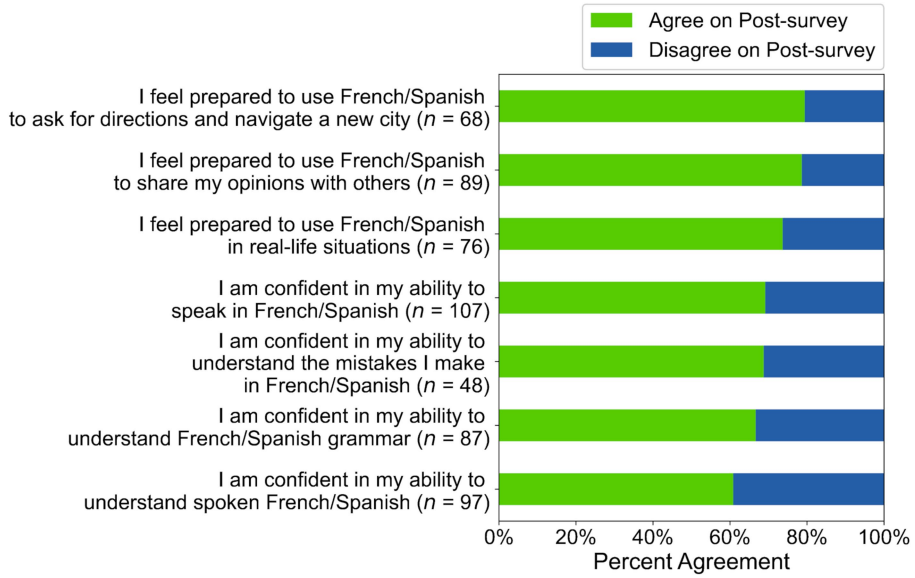

FIGURE 9
Post-survey agreement for participants who disagreed on the pre-survey in Study 2.

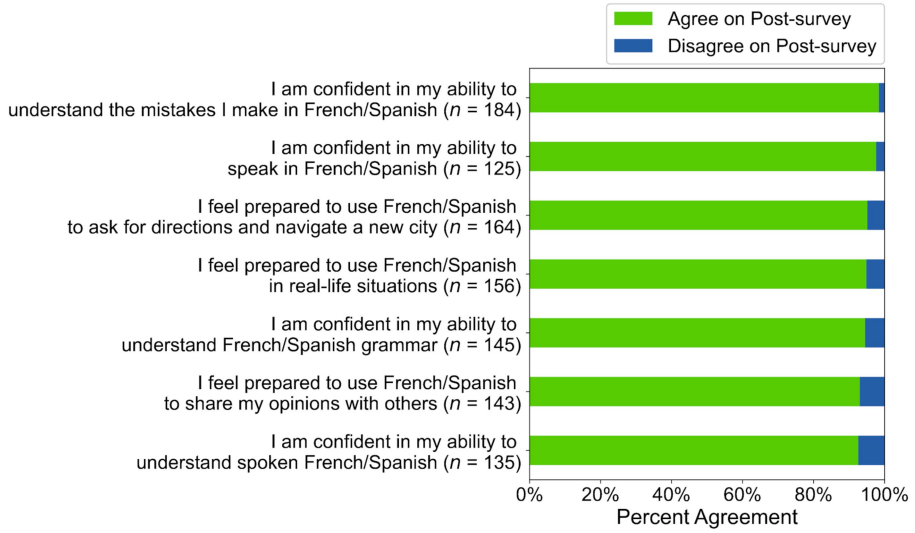

FIGURE 10
Post-survey agreement for participants who agreed on the pre-survey in Study 2.

self-efficacy for language skills, after using Duolingo AI-based features for 1 month. Learners' perceptions of the features' efficacy and self-reported transfer of learning outside the app were also analyzed. Study 2 evaluated the same outcomes in learners who were given 1 month of complimentary access to the AI-based features.

7.1 Increases in learners' self-efficacy

Participants who engaged in learning experiences enabled by the AI-based features showed significant increases in their self-efficacy, especially when the AI-based features were novel. In Study 1, self-efficacy for using French/Spanish in real-life situations increased significantly. In Study 2, learners showed a significant increase in self-efficacy for using French/Spanish in real-life situations, as well as for the specific situations of navigating a new city and sharing opinions with others. The learners in Study 2 also showed significant increases in self-efficacy for speaking in French/Spanish, understanding French/Spanish grammar, and understanding mistakes in French/Spanish.

These results represent the first evidence that language learning self-efficacy increases significantly after app-based learning experiences that use generative AI, consistent with learners' expectation that generative AI in language apps can help them with

TABLE 3 Full regression output for Study 2 (N = 232).

Statement	Regression formula	Term	Coeff	SE	z	p value
"I feel prepared to use French/Spanish in real-life situations"	Post-survey ~ Intercept + Pre-survey + RP + EMA	Intercept	**1.04**	**0.27**	**3.88**	**p < 0.001**
		Pre-survey	**1.95**	**0.46**	**4.26**	**p < 0.001**
		RP	0.02	0.02	1.23	0.27
		EMA	−0.01	0.01	−1.23	0.27
"I feel prepared to use French/Spanish to ask for directions and navigate a new city"	Post-survey ~ Intercept + Pre-survey + RP + EMA	Intercept	**1.66**	**0.39**	**4.27**	**p < 0.001**
		Pre-survey	**1.74**	**0.49**	**3.58**	**p < 0.01**
		RP	0.08	0.04	2.21	0.06
		EMA	−0.01	0.01	−1.35	0.26
"I feel prepared to use French/Spanish to share my opinions with others"	Post-survey ~ Intercept + Pre-survey + RP + EMA	Intercept	**1.33**	**0.27**	**5.01**	**p < 0.001**
		Pre-survey	**1.29**	**0.42**	**3.05**	**p < 0.01**
		RP	0.02	0.02	1.23	0.27
		EMA	−0.003	0.01	−0.66	0.53
"I'm confident in my ability to understand spoken French/Spanish"	Post-survey ~ Intercept + Pre-survey + RP + EMA	Intercept	0.45	0.22	2.05	0.07
		Pre-survey	**2.19**	**0.40**	**5.48**	**p < 0.001**
		RP	0.03	0.02	2.10	0.07
		EMA	−0.01	0.01	−1.13	0.30
"I'm confident in my ability to speak in French/Spanish"	Post-survey ~ Intercept + Pre-survey + RP + EMA + EMA^2	Intercept	**0.94**	**0.23**	**4.04**	**p < 0.001**
		Pre-survey	**2.86**	**0.62**	**4.58**	**p < 0.001**
		RP	0.02	0.02	1.44	0.23
		EMA	−0.02	0.01	−1.67	0.15
		EMA^2	0.0001	0.00007	1.32	0.26
"I'm confident in my ability to understand French/ Spanish grammar"	Post-survey ~ Intercept + Pre-survey + RP + EMA	Intercept	**0.70**	**0.24**	**2.99**	**p < 0.01**
		Pre-survey	**2.19**	**0.43**	**5.06**	**p < 0.001**
		RP	0.03	0.02	1.67	0.15
		EMA	0.0006	0.01	0.12	0.91
"I'm confident in my ability to understand the mistakes I make in French/Spanish"	Post-survey ~ Intercept + Pre-survey + RP + EMA	Intercept	**0.81**	**0.32**	**2.53**	**p < 0.05**
		Pre-survey	**3.35**	**0.67**	**5.03**	**p < 0.001**
		RP	−0.01	0.01	−0.82	0.44
		EMA	0.01	0.01	0.83	0.44

Coeff, Ceofficient; SE, Coefficient Standard Error; RP, mean-centered number of Roleplays completed; EMA, mean-centered number of times the Explain My Answer feature was shown; EMA2, EMA squared then mean-centered. Values for Coefficients with $p < 0.05$ are bolded.

personalized conversation practice and error correction (Yuen and Schlote, 2024). The findings in this study are consistent with evidence that technology-assisted learning experiences prior to ChatGPT support language learning self-efficacy (Zhang, 2022), including studies in China (Liu, 2020), Iran (Babakhani and Tabatabaee-Yazdi, 2023; Dong et al., 2022), and the US (Zheng et al., 2009). More broadly, these findings build on previous international research showing that language learning self-efficacy can be increased through interventions with communication-focused tasks, constructive feedback that highlights learner success, and explicit teaching (Raoofi et al., 2012; Chen, 2022; Goetze and Driver, 2022). Taken together with the present results, this literature suggests that other AI-enabled learning experiences that effectively leverage communication-focused tasks, constructive feedback, and explicit teaching should also be able to increase learners' self-efficacy. More research with other AI-based tools is needed to test this hypothesis.

Since many app learners are studying a language for real-world communication (Kittredge and Peters, 2023), these AI-based features may boost self-efficacy for skills that are directly relevant to learners' goals. The findings also answer calls for more research on specific language skill self-efficacy in non-English languages (Goetze and Driver, 2022). The increased self-efficacy for speaking in Study 2 is important, since speaking self-efficacy correlates strongly with performance (Goetze and Driver, 2022), and preventing disappointment in oral proficiency may protect against language learning demotivation (Albalawi and Al-Hoorie, 2021). Although these studies did not measure language learning outcomes, self-efficacy is correlated with learning outcomes in a variety of language learning contexts (Goetze and Driver, 2022) including Duolingo (Jiang et al., 2024a), and there is evidence that language learning apps are effective at teaching language skills (Tommerdahl et al., 2024). Taken together, this evidence suggests that learners in the present

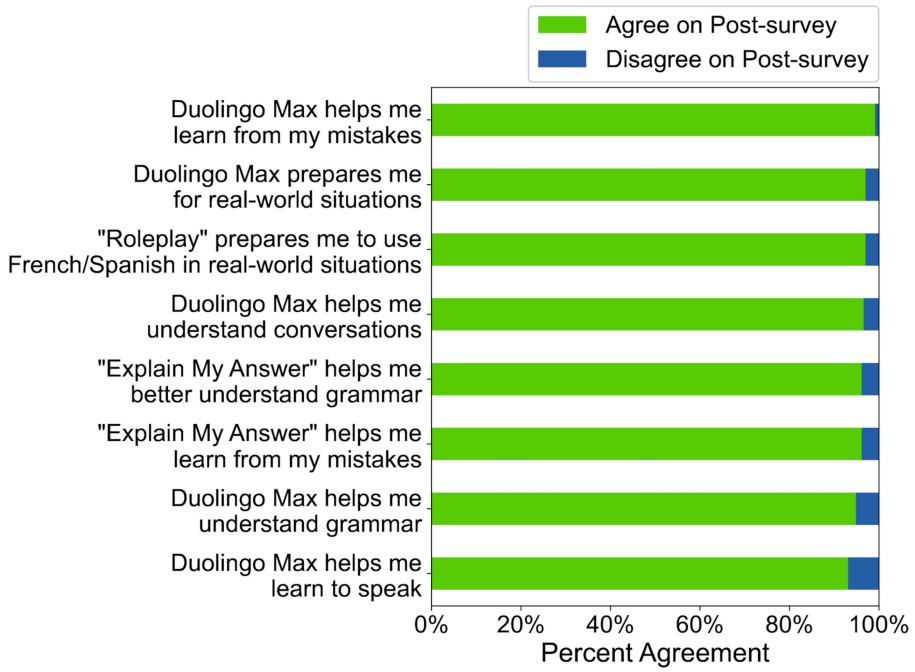

FIGURE 11
Agreement rates for 8 statements on the post-survey in Study 2 (*N* = 232). Note: "Duolingo Max" was the name of the higher subscription tier at the time of the study.

TABLE 4 Summary of AI-generated, human verified themes in Study 2 participants' descriptions of using what they learned outside of Duolingo (*N* = 121).

Theme	Examples
Travel	Asking for directions, ordering food, and conversing with locals.
Casual conversations	With friends, family, and neighbors who speak the language they are learning. Sometimes via social media or texting.
Workplace communication	With coworkers, clients, and during work-related travel.
Cultural engagement	Attending festivals, watching movies, listening to music, and reading articles or books.
Educational support and settings	Assisting students with homework, teaching children, or practicing with a partner who is also learning the language.
Daily errands and services	Grocery shopping, ordering at restaurants, and interacting with service providers who speak the language.
Family interaction	With spouses, children, and relatives. Often with the goal of better communication with family who are native speakers.

studies may have improved their language skills, in addition to their self-efficacy.

The intervention in the present studies was relatively intense in its weekly usage requirement, but the total duration was short compared to previous interventions that lasted a semester or more (Chen, 2022). Future research is needed to determine whether self-efficacy would increase, plateau, or possibly decline with longer exposure to learning experiences enabled by AI-based features. Also, learners in Study 1 were restricted to English-speaking countries. While learners in Study 2 did not have the same location restriction, they were still limited to English speakers learning Spanish or French. More research on the development of self-efficacy with generative AI is needed with a variety of language learners and geographical contexts.

7.2 Perceived efficacy of AI-based features

Across Studies 1 and 2, 91–97% of learners agreed that the AI-based features helped them learn to speak and understand conversations, similar to another study conducted in Benin in which learners reported that generative AI improved their English speaking skills (Toboula, 2023). In Studies 1 and 2, 90–97% of learners agreed that the AI-based features, including the Roleplay feature, prepared them for real-world situations, and 86–99% agreed that the AI-based features, including the Explain My Answer feature, helped them understand grammar and learn from their mistakes. These results imply that learners noticed improvements in their skills during the study.

7.3 Self-reported transfer of learning

The majority of learners in both studies reported using what they learned with the AI-based features outside the app, in similar contexts across studies: Travel, work, education, services and retail, interactions with family and friends, casual conversations with native speakers, communication via technology, and media use. Taken

together with the high perceived efficacy of AI-based features, these findings suggest that many learners noticed transfer of learning from the app to real life. Such perceptions could play a role in supporting actual skill development, because people's judgments of their own learning can influence their approach to future learning (Metcalfe, 2009).

While learners attributed learning transfer to the AI-based features, most of the course content learners studied was the same as in the free version of the app, and this content could also have contributed to learning transfer. Future research could investigate whether transfer of learning is observed in app learning experiences that do not rely heavily on generative AI.

7.4 Comparison of studies 1 and 2

Although we do not directly compare the results of Studies 1 and 2 because they represent different learner populations, we note some important differences. In Study 2 there were significant increases in several different types of self-efficacy, and perceived efficacy of the AI-based features ranged from 93 to 99%. In Study 1, just one type of self-efficacy increased significantly, and AI-based feature efficacy perception ranged from 86 to 96%.

While it is difficult to attribute these differences to any one factor, there are several possible sources. Learners in Study 2 could have experienced a novelty effect, viewing their experience with AI-based features more positively because they were novel. This is in line with research demonstrating enhanced interest in language learning when interactions with chatbots are novel (Fryer et al., 2017; Fryer et al., 2019), and research showing more positive attitudes towards technologies that are perceived as more novel (Wells et al., 2010). These results are also broadly consistent with research demonstrating that novelty enhances attention, memory, learning, and perceived value (Barto et al., 2013; Kagan, 2009; Blut et al., 2023; Sánchez-Fernández and Iniesta-Bonillo, 2007; Leroi-Werelds, 2019; Wells et al., 2010). Furthermore, learners in Study 2 may have valued the AI-based features more positively because they did not pay for them (Blut et al., 2023; Lee et al., 2020; Wang et al., 2013). Study 2's larger sample size may also have made it easier to detect effects. The non-paying learners who completed Study 2 were also slightly younger (32% 18–34 years old, 68% 35+ years old) than the subscribers who completed Study 1 (19% 18–34 years old, 81% 35+ years old), and their location was not limited to English-speaking countries. The different attrition rates in the studies may also suggest underlying differences in income or commitment to language learning between subscribers and non-paying learners.

7.5 Ethical implications

Concerns have been raised about the ethical implications of using generative AI for educational and research purposes (Bond et al., 2024). Biases against specific demographic groups in text generated by large language models, as well as lack of transparency about data privacy, are well-known risks (Landers and Behrend, 2023). More recently, studies have shown that generative AI can even harm subsequent learning when it includes errors or is used as a crutch during the learning process (Bastani et al., 2024).

Several factors mitigate these concerns in the present studies. Individual participants' data collected via study surveys were not shared with anyone outside the company, and participants were informed of this. Duolingo's privacy policy regarding in-app data is also publicly available (Duolingo, 2024). The prompts that generated text live in Duolingo's AI-based features were specially designed to avoid bias, errors, and inappropriate or offensive language (Duolingo Team, 2023). Finally, although some of the Duolingo AI-based features provided translation hints and helpful phrase suggestions, learners still needed to assemble phrases into complete responses during conversation activities. Future research is needed to investigate the degree to which generative AI-based assistance during learning experiences enhances or hinders language skill development.

7.6 Limitations of the current studies

While these studies' results are encouraging, they have several limitations. The study design did not include a control group (i.e., learners who did not use the AI-based features), so it is possible that self-efficacy increases would have occurred without access to the AI-based features, or were even unrelated to Duolingo. Also, since most learners reported using their learning outside the app, it is possible (although unlikely) that these practice opportunities alone led to increased self-efficacy. Future studies aiming to identify causality should include a control group.

Furthermore, these studies rely on self-reported data. Although assessing self-efficacy via questionnaires is common practice in second language acquisition research (Goetze and Driver, 2022), in any self-reported data there is the potential of bias. People's judgments of learning are influenced by ease of processing, which can sometimes be unrelated to or even negatively correlated with learning (Carpenter and Geller, 2020; Deslauriers et al., 2019). For instance, the Roleplay feature could have made learners *feel* like they could speak easily without significantly increasing speaking skills. However, the high rate of participants' self-reported transfer of learning outside of the app provides a check on the accuracy of their self-efficacy estimates, because learners who use the language in daily life get immediate feedback on their skill level. While this triangulation of varied self-report data lends credibility to participants' self-efficacy estimates, future research should investigate whether such AI-based features empirically improve learning.

Finally, these results do not include learners who did not meet the study's app usage requirements, perhaps due to lower perceived efficacy of the AI-based features, or lesser motivation or language ability. Also, the learners in this study are English speakers learning Spanish or French, and the results may or may not generalize to other populations. Future research on the use of AI-based features to support self-efficacy should be conducted with a greater diversity of language learners.

8 Conclusion

In two studies, learners used the Duolingo language learning app for 1 month, with access to two features that used generative AI. These features prepared learners for real-world conversations and provided on-demand explanations, using pedagogical approaches that

enhanced self-efficacy in previous classroom interventions. Changes in learners' self-efficacy were assessed, as well as learners' perceived efficacy of the AI-based features and self-reported transfer of learning from the app to real life.

In Study 1, learners who had already been using the AI-based features as part of a paid subscription tier showed a significant increase in their feelings of preparedness to use French/Spanish in real-life situations. In Study 2, learners granted complimentary access to the AI-based features showed significant increases in self-efficacy for a variety of skills, including feeling prepared to use French/Spanish in several real-life situations, and feeling confident in their ability to speak in French/Spanish, understand French/Spanish grammar, and understand the mistakes they made. A large majority of learners in both studies agreed that the AI-based features were effective in helping them acquire a variety of language skills, suggesting that learners experienced improvements in their abilities during the study. Consistent with this interpretation, the majority of learners said they had used what they learned outside the app, and gave examples from a variety of real-world contexts.

These findings are the first evidence that language learning self-efficacy increases in learning experiences that use generative AI. These results are consistent with previous intervention studies that enhanced language learning self-efficacy with communication-focused tasks, constructive feedback that highlights learners' success, and explicit teaching of language learning strategies.

Data availability statement

The datasets presented in this article are not readily available because the datasets are the property of Duolingo and cannot currently be released. Requests to access the datasets should be directed to learning-sciences@duolingo.com.

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

AK: Writing – original draft, Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Visualization, Writing – review & editing. EH: Data curation, Formal analysis, Methodology, Visualization, Writing – original draft, Writing – review & editing. BR: Conceptualization, Data curation, Investigation, Methodology, Project administration, Writing – review & editing. DD: Data curation, Formal analysis, Methodology, Writing – review & editing. CF: Conceptualization, Funding acquisition, Methodology, Supervision, Writing – review & editing. XJ: Supervision, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This study received funding from Duolingo. The funder had the following involvement with the study: approval of research study design, provision of funding for participant compensation and software licenses.

Acknowledgments

Many thanks to: Dr. Lucy Skidmore for help with collecting data, analysis, and formatting the report; Aslı Yurtsever for help with formatting figures; Kai Li for help with granting participants subscriptions; Leah Goldman for help with accessing and analyzing participant data; Sarika Patel for input on survey item wording; Dr. Bozena Pajak for feedback during various stages of this project; Dr. Luke Plonsky and the reviewers for comments on this manuscript.

Conflict of interest

AK, EH, BR, DD, CF, and XJ were employed by Duolingo.

Publisher's note

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2025.1499497/full#supplementary-material

References

Abdelhalim, S. M. (2024). From traditional writing to digital multimodal composing: promoting high school EFL students' writing self-regulation and self-efficacy. *Comput. Assist. Lang. Learn.* 30:148. doi: 10.1080/09588221.2024.2322148

Albalawi, F. H., and Al-Hoorie, A. H. (2021). From demotivation to remotivation: a mixed-methods investigation. *SAGE Open* 11:101. doi: 10.1177/21582440211041101

Babakhani, A., and Tabatabaee-Yazdi, M. (2023). The power of gamification on Iranian EFL learners' self-efficacy. *J. New Adv. English Lang. Teach. Appl. Linguist.* 5, 1118–1129. doi: 10.22034/jeltal.2023.5.1.4

Bai, Y. (2024). A mixed methods investigation of Mobile-based language learning on EFL students' listening, speaking, foreign language enjoyment, and anxiety. *SAGE Open* 14, 1–19. doi: 10.1177/21582440241255554

Bandura, A. (1994). "Self-efficacy" in Encyclopedia of human behavior. ed. V. S. Ramachaudran (London: Academic Press), 71–81.

Barto, A., Mirolli, M., and Baldassarre, G. (2013). Novelty or surprise? *Front. Psychol.* 4:907. doi: 10.3389/fpsyg.2013.00907

Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, O., and Mariman, R., (2024) Generative AI can harm learning (July 15, 2024). The Wharton School Research Paper, Available at: https://ssrn.com/abstract=4895486 (Accessed January 13, 2025).

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological) 57,* 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Bernacki, M. L., Nokes-Malach, T. J., and Aleven, V. (2015). Examining self-efficacy during learning: variability and relations to behavior, performance, and learning. *Metacogn. Learn.* 10, 99–117. doi: 10.1007/s11409-014-9127-x

Blut, M., Chaney, D., Lunardo, R., Mencarelli, R., and Grewal, D. (2023). Customer perceived value: A comprehensive meta-analysis. *J. Serv. Res.* 27, 501–524. doi: 10.1177/10946705231222295

Bond, M., Khosravi, H., De Laat, M., Bergdahl, N., Negrea, V., Oxley, E., et al. (2024). A meta systematic review of artificial intelligence in higher education: a call for increased ethics, collaboration, and rigour. *Int. J. Educ. Technol. High. Educ.* 21:4. doi: 10.1186/s41239-023-00436-z

Carpenter, S. K., and Geller, J. (2020). Is a picture really worth a thousand words? Evaluating contributions of fluency and analytic processing in metacognitive judgements for pictures in foreign language vocabulary learning. *Q. J. Exp. Psychol.* 73, 211–224. doi: 10.1177/1747021819879416

Chen, J. (2022). The effectiveness of self-regulated learning (SRL) interventions on L2 learning achievement, strategy employment and self-efficacy: A meta-analytic study. *Front. Psychol.* 13:101. doi: 10.3389/fpsyg.2022.1021101

Council of Europe (2001). Common European framework of references for languages: learning, teaching, assessment. Cambridge: Cambridge University Press.

Deslauriers, L., McCarty, L. S., Miller, K., Callaghan, K., and Kestin, G. (2019). Measuring actual learning versus feeling of learning in response to being actively engaged in the classroom. *Proc. Natl. Acad. Sci.* 116, 19251–19257. doi: 10.1073/pnas.1821936116

Dong, L., Jamal Mohammed, S., Ibrahim, A. A.-A., and Rezai, A. (2022). Fostering EFL learners' motivation, anxiety, and self-efficacy through computer-assisted language learning-and mobile-assisted language learning-based instructions. *Front. Psychol.* 13:899557. doi: 10.3389/fpsyg.2022.899557

Du, J., and Daniel, B. K. (2024). Transforming language education: A systematic review of AI-powered chatbots in EFL speaking practice. *Computers and education. Artif. Intell.* 6:100230. doi: 10.1016/j.caeai.2024.100230

Duolingo (2024). Privacy policy. https://www.duolingo.com/privacy (Accessed March 1, 2024).

Duolingo Team (2023). Introducing Duolingo max, a learning experience powered by GPT-4. Duolingo blog. Available at: https://blog.duolingo.com/duolingo-max/ (Accessed June 25, 2024).

Fan, X. (2023). Accelerated English teaching methods: the role of digital technology. *J. Psycholinguist. Res.* 52, 1545–1558. doi: 10.1007/s10936-023-09961-4

Freeman, C., Kittredge, A., Wilson, H., and Pajak, B. (2023). The Duolingo method for app-based teaching and learning [white paper]. Available at: https://duolingo-papers.s3.amazonaws.com/reports/Duolingo_whitepaper_duolingo_method_2023.pdf (Accessed June 25, 2024).

Fryer, L. K., Ainley, M., Thompson, A., Gibson, A., and Sherlock, Z. (2017). Stimulating and sustaining interest in a language course: an experimental comparison of Chatbot and human task partners. *Comput. Hum. Behav.* 75, 461–468. doi: 10.1016/j.chb.2017.05.045

Fryer, L. K., Nakao, K., and Thompson, A. (2019). Chatbot learning partners: connecting learning experiences, interest and competence. *Comput. Hum. Behav.* 93, 279–289. doi: 10.1016/j.chb.2018.12.023

García Botero, G., Botero Restrepo, M. A., Zhu, C., and Questier, F. (2021). Complementing in-class language learning with voluntary out-of-class MALL. Does training in self-regulation and scaffolding make a difference? *Comput. Assist. Lang. Learn.* 34, 1013–1039. doi: 10.1080/09588221.2019.1650780

Glean (2024). Glean [AI-powered work assistant]. Available at: https://www.glean.com/

Godwin-Jones, R. (2024). Distributed agency in second language learning and teaching through generative AI. *Lang. Learn. Technol.* 28, 5–31.

Goetze, J., and Driver, M. (2022). Is learning really just believing? A meta-analysis of self-efficacy and achievement in SLA. *Stud. Sec. Lang. Learn. Teach.* 12, 233–259. doi: 10.14746/ssllt.2022.12.2.4

Graham, S. (2022). Self-efficacy and language learning-what it is and what it isn't. *Lang. Learn. J.* 50, 186–207. doi: 10.1080/09571736.2022.2045679

Han, Z. (2024). ChatGPT in and for second language acquisition: A call for systematic research. *Stud. Second. Lang. Acquis.* 46, 301–306. doi: 10.1017/S0272263124000111

Harpe, S. E. (2015). How to analyze Likert and other rating scale data. *Curr Pharm Teach Learn* 7, 836–850. doi: 10.1016/j.cptl.2015.08.001

Harper, D., Bowles, A. R., Amer, L., Pandža, N. B., and Linck, J. A. (2021). Improving outcomes for English learners through technology: A randomized controlled trial. *Aera Open* 7, 1–20. doi: 10.1177/23328584211025528

Honicke, T., and Broadbent, J. (2016). The influence of academic self-efficacy on academic performance: A systematic review. *Educ. Res. Rev.* 17, 63–84. doi: 10.1016/j.edurev.2015.11.002

Jiang, X., Hopman, E., Reuveni, B., and Kittredge, A. (2024a). Duolingo path meets expectations for proficiency outcomes [white paper]. Available at: https://duolingo-papers.s3.amazonaws.com/reports/Duolingo_whitepaper_language_read_listen_write_speak_2024.pdf (Accessed June 25, 2024).

Jiang, X., Peters, R., Plonsky, L., and Pajak, B. (2024b). The effectiveness of Duolingo English courses in developing reading and listening proficiency. *CALICO J.* doi: 10.1558/cj.26704

Jiang, X., Rollinson, J., Plonsky, L., Gustafson, E., and Pajak, B. (2021). Evaluating the reading and listening outcomes of beginning-level Duolingo courses. *Foreign Lang. Ann.* 54, 974–1002. doi: 10.1111/flan.12600

Kagan, J. (2009). Categories of novelty and states of uncertainty. *Rev. Gen. Psychol.* 13, 290–301. doi: 10.1037/a0017142

Kessler, M., Loewen, S., and Gönülal, T. (2023). Mobile-assisted language learning with Babbel and Duolingo: comparing L2 learning gains and user experience. *Comput. Assist. Lang. Learn.* 1, 1–25. doi: 10.1080/09588221.2023.2215294

Kittredge, A., and Peters, R. (2023). Special report: which country studies English the most? Duolingo blog. Available at: https://blog.duolingo.com/which-country-studies-english-the-most/ (Accessed June 25, 2024).

Landers, R. N., and Behrend, T. S. (2023). Auditing the AI auditors: a framework for evaluating fairness and bias in high stakes AI predictive models. *Am. Psychol.* 78, 36–49. doi: 10.1037/amp0000972

Law, L. (2024). Application of generative artificial intelligence (GenAI) in language teaching and learning: A scoping literature review. *Comput. Educ. Open* 6:100174. doi: 10.1016/j.caeo.2024.100174

Lee, Y. S., Jung, W. M., Bingel, U., and Chae, Y. (2020). The context of values in pain control: understanding the price effect in placebo analgesia. *J. Pain* 21, 781–789. doi: 10.1016/j.jpain.2019.11.005

Leeming, P. (2017). A longitudinal investigation into English speaking self-efficacy in a Japanese language classroom. *Asia-Pac. J. Second Foreign Lang. Educ.* 2, 1–18. doi: 10.1186/s40862-017-0035-x

Leroi-Werelds, S. (2019). An update on customer value: state of the art, revised typology, and research agenda. *J. Serv. Manag.* 30, 650–680. doi: 10.1108/JOSM-03-2019-0074

Li, J., Wang, C., Zhao, Y., and Li, Y. (2023). Boosting learners' confidence in learning English: can self-efficacy-based intervention make a difference? *Tesol Q.* 58, 1518–1547. doi: 10.1002/tesq.3292

Liu, M. (2020). The effect of mobile learning on students' reading self-efficacy: A case study of the app "English Liulishuo". *Engl. Lang. Teach.* 13, 91–101. doi: 10.5539/elt.v13n12p91

Marzilli Ericson, K. M., and Fuster, A. (2014). The endowment effect. *Annu. Rev. Econ* 6, 555–579. doi: 10.1146/annurev-economics-080213-041320

Metcalfe, J. (2009). Metacognitive judgments and control of study. *Curr. Dir. Psychol. Sci.* 18, 159–163. doi: 10.1111/j.1467-8721.2009.01628.x

Milliner, B., and Dimoski, B. (2024). The effects of a metacognitive intervention on lower-proficiency EFL learners' listening comprehension and listening self-efficacy. *Lang. Teach. Res.* 28, 679–713. doi: 10.1177/13621688211004646

Mills, N. (2009). A guide du Routard simulation: increasing self-efficacy in the standards through project-based learning. *Foreign Lang. Ann.* 42, 607–639. doi: 10.1111/j.1944-9720.2009.01046.x

Moorhouse, B. L., and Kohnke, L. (2024). The effects of generative AI on initial language teacher education: the perspectives of teacher educators. *System* 122:103290. doi: 10.1016/j.system.2024.103290

Morgan, D. L. (2023). Exploring the use of artificial intelligence for qualitative data analysis: the case of ChatGPT. *Int J Qual Methods* 22:248. doi: 10.1177/16094069231211248

Nguyen, A. T., Nguyen, T. T., Le, T. T., Phuong, H. Y., Pham, T. T., Huynh, T. A. T., et al. (2023). Effects of Memrise on Vietnamese EFL Students' vocabulary: A case study at a College in a Rural Area. *Electr. J. Learn* 21, 450–460. doi: 10.34190/ejel.21.5.3066

OpenAI (2022). Text-embedding-ada-002 [Large language model]. Available at: https://platform.openai.com/docs/api-reference/embeddings (Accessed March 14, 2024).

OpenAI (2023a). ChatGPT-3.5-turbo [Large language model]. Available at: https://platform.openai.com/docs/guides/fine-tuning (Accessed March 14, 2024).

OpenAI (2023b). ChatGPT-4 [Large language model]. Available at: https://openai.com/gpt-4 (Accessed February 28, 2024).

Pichugin, V., Panfilov, A., and Volkova, E. (2023). Applications with memory load for lexical activation in foreign language learning. *Front. Educ.* 8:1278541. doi: 10.3389/feduc.2023.1278541

Qiao, H., and Zhao, A. (2023). Artificial intelligence-based language learning: illuminating the impact on speaking skills and self-regulation in Chinese EFL context. *Front. Psychol.* 14:1255594. doi: 10.3389/fpsyg.2023.1255594

Rachels, J. R., and Rockinson-Szapkiw, A. J. (2018). The effects of a mobile gamification app on elementary students' Spanish achievement and self-efficacy. *Comput. Assist. Lang. Learn.* 31, 72–89. doi: 10.1080/09588221.2017.1382536

Raoofi, S., Tan, B. H., and Chan, S. H. (2012). Self-efficacy in second/foreign language learning contexts. *English. Lang. Teach.* 5:60. doi: 10.5539/elt.v5n11p60

Royston, P., Altman, D. G., and Sauerbrei, W. (2006). Dichotomizing continuous predictors in multiple regression: a bad idea. *Stat. Med.* 25, 127–141. doi: 10.1002/sim.2331

Sánchez-Fernández, R., and Iniesta-Bonillo, M. Á. (2007). The concept of perceived value: a systematic review of the research. *Mark. Theory* 7, 427–451. doi: 10.1177/1470593107083165

Schneider, M., and Preckel, F. (2017). Variables associated with achievement in higher education: A systematic review of meta-analyses. *Psychol. Bull.* 143, 565–600. doi: 10.1037/bul0000098

Shamshiri, F., Shafiee, S., and Rahimi, F. (2023). The effects of computer-assisted language learning (CALL) and different interaction patterns on vocabulary development of EFL learners. *J. Lang. Educ* 9, 110–127. doi: 10.17323/jle.2023.12093

Shortt, M., Tilak, S., Kuznetcova, I., Martens, B., and Akinkuolie, B. (2023). Gamification in mobile-assisted language learning: A systematic review of Duolingo literature from public release of 2012 to early 2020. *Comput. Assist. Lang. Learn.* 36, 517–554. doi: 10.1080/09588221.2021.1933540

Statista. (2024). Leading language learning apps worldwide in January 2024, by downloads. Available at: https://www.statista.com/statistics/1239522/top-language-learning-apps-downloads/ (Accessed June 25, 2024).

Talsma, K., Schüz, B., Schwarzer, R., and Norris, K. (2018). I believe, therefore I achieve (and vice versa): A meta-analytic cross-lagged panel analysis of self-efficacy and academic performance. *Learn. Individ. Differ.* 61, 136–150. doi: 10.1016/j.lindif.2017.11.015

Toboula, C. M. Z. (2023). Enhancing post-pandemic EFL education by leveraging immersive, NLP-driven, AI-based tools that promote collaboration and interactivity within an educational approach. *Int. J. Cybern. Inform.* 12, 171–193. doi: 10.5121/ijci.2023.120213

Tommerdahl, J. M., Dragonflame, C. S., and Olsen, A. A. (2024). A systematic review examining the efficacy of commercially available foreign language learning mobile apps. *Comput. Assist. Lang. Learn.* 37, 333–362. doi: 10.1080/09588221.2022.2035401

Vu, T., Magis-Weinberg, L., Jansen, B. R., van Atteveldt, N., Janssen, T. W. P., Lee, N. C., et al. (2022). Motivation-achievement cycles in learning: A literature review and research agenda. *Educ. Psychol. Rev.* 34, 39–71. doi: 10.1007/s10648-021-09616-7

Wang, L. C., Lam, E. T. C., and Xiao, C. (2023). The effectiveness of using Memrise application to learn Chinese characters by American middle school students - A pilot study. *Int. J. Technol. Educ.* 6, 583–592. doi: 10.46328/ijte.423

Wang, T., Oh, L.-B., Wang, K., and Yuan, Y. (2013). User adoption and purchasing intention after free trial: an empirical study of mobile newspapers. *IseB* 11, 189–210. doi: 10.1007/s10257-012-0197-5

Wang, C., and Sun, T. (2020). Relationship between self-efficacy and language proficiency: a meta-analysis. *System* 95:102366. doi: 10.1016/j.system.2020.102366

Wells, J. D., Campbell, D. E., Valacich, J. S., and Featherman, M. (2010). The effect of perceived novelty on the adoption of information technology innovations: a risk/reward perspective. *Decis. Sci.* 41, 813–843. doi: 10.1111/j.1540-5915.2010.00292.x

Wigfield, A., and Eccles, J. S. (2000). Expectancy–value theory of achievement motivation. *Contemp. Educ. Psychol.* 25, 68–81. doi: 10.1006/ceps.1999.1015

Wittmann, B. C., Daw, N. D., Seymour, B., and Dolan, R. J. (2008). Striatal activity underlies novelty-based choice in humans. *Neuron* 58, 967–973. doi: 10.1016/j.neuron.2008.04.027

Xiao, F., Zhao, P., Sha, H., Yang, D., and Warschauer, M. (2023). Conversational agents in language learning. *J. China Comput. Assist. Lang. Learn.* 4, 300–325. doi: 10.1515/jccall-2022-0032

Xu, M., Wang, C., Chen, X., Sun, T., and Ma, X. (2022). Improving self-efficacy beliefs and English language proficiency through a summer intensive program. *System* 107:102797. doi: 10.1016/j.system.2022.102797

Yang, H., and Lian, Z. (2023). Ideal L2 self, self-efficacy, and pragmatic production: the mediating role of willingness to communicate in learning English as a foreign language. *Behav. Sci.* 13:597. doi: 10.3390/bs13070597

Young Kyo, O. H. (2022). The growth trajectories of L2 self-efficacy and its effects on L2 learning: using a curve-of-factors model. *Appl. Linguis.* 43, 147–167. doi: 10.1093/applin/amab005

Yuen, C. L., and Schlote, N. (2024). Learner experiences of mobile apps and artificial intelligence to support additional language learning in education. *J. Educ. Technol. Syst.* 52, 507–525. doi: 10.1177/00472395241238693

Zhang, Y. (2022). The effect of educational technology on EFL learners' self-efficacy. *Front. Psychol.* 13:881301. doi: 10.3389/fpsyg.2022.881301

Zheng, D., Young, M. F., Brewer, R. A., and Wagner, M. (2009). Attitude and self-efficacy change: English language learning in virtual worlds. *Calico J.* 27, 205–231. doi: 10.11139/cj.27.1.205-231

Zimmerman, B. J. (2000). Self-efficacy: an essential motive to learn. *Contemp. Educ. Psychol.* 25, 82–91. doi: 10.1006/ceps.1999.1016