Evaluating LLM-Generated Lessons from the Language Learning Students' Perspective: A Short Case Study on Duolingo

Carlos Rafael Catalan
Samsung R&D Institute Philippines
Manila, Philippines
c.catalan@samsung.com

Patricia Nicole Monderin
Samsung R&D Institute Philippines
Manila, Philippines
p.monderin@samsung.com

Lheane Marie Dizon
Samsung R&D Institute Philippines
Manila, Philippines
lm.dizon@samsung.com

Gap Estrella
Samsung R&D Institute Philippines
Manila, Philippines
pg.estrella@samsung.com

Raymund John Sarmiento
Samsung R&D Institute Philippines
Manila, Philippines
rj.sarmiento@samsung.com

Marie Antoinette Patalagsa
Samsung R&D Institute Philippines
Manila, Philippines
m.patalagsa@samsung.com

Abstract

Popular language learning applications such as Duolingo use large language models (LLMs) to generate lessons for its users. Most lessons focus on general real-world scenarios such as greetings, ordering food, or asking directions, with limited support for profession-specific contexts. This gap can hinder learners from achieving professional-level fluency, which we define as the ability to communicate comfortably various work-related and domain-specific information in the target language. We surveyed five employees from a multinational company in the Philippines on their experiences with Duolingo. Results show that respondents encountered general scenarios more frequently than work-related ones, and that the former are relatable and effective in building foundational grammar, vocabulary, and cultural knowledge. The latter helps bridge the gap toward professional fluency as it contains domain-specific vocabulary. Each participant suggested lesson scenarios that diverge in contexts when analyzed in aggregate. With this understanding, we propose that language learning applications should generate lessons that adapt to an individual's needs through personalized, domain-specific lesson scenarios while maintaining foundational support through general, relatable lesson scenarios.

Keywords

Large Language Models, Language Learning, Intelligent Tutoring Systems, User-Centered Evaluation

ACM Reference Format:
Carlos Rafael Catalan, Patricia Nicole Monderin, Lheane Marie Dizon, Gap Estrella, Raymund John Sarmiento, and Marie Antoinette Patalagsa. 2026. Evaluating LLM-Generated Lessons from the Language Learning Students' Perspective: A Short Case Study on Duolingo. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (3rd HEAL Workshop - CHI)*. ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

3rd HEAL Workshop - CHI, Barcelona, Spain

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Large Language Models (LLMs) show exceptional capabilities in educational use-cases, such as generating lessons for students [10, 11, 21]. In the domain of language learning, this technology has made language acquisition applications such as Duolingo [13] a popular tool for users to acquire fluency in another language [2, 15]. In line with traditional language learning settings, generated lessons typically depict real-world scenarios in the target language to immerse the student to help them gain proficiency and fluency[5].

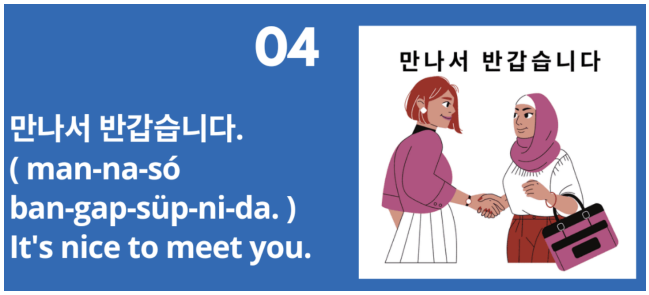

Figure 1: A scenario showing a greeting in Korean. Presenting a real-world scenario is typical of language learning resources, as it is able to immerse the learner in the target language. (Image was grabbed from https://www.topikguide.com/korean-greetings/)

To accommodate students of diverse backgrounds, these lessons often depict general scenarios, such as greetings, ordering food, and asking for directions. However, very rarely do these applications provide scenarios that are catered to students' unique professional or office settings such as communicating technical project specifications to stakeholders, or negotiating deadlines with project managers. This presents a gap for students working in multinational and multilingual environments. Here, a student must not only be able to fluently communicate in the target language general scenarios such as greetings, but also more technical, domain-specific discussions about their work. The latter may not always be covered by the current general scenarios which these applications generate.

In this workshop contribution, we conduct an exploratory study and provide a preliminary evaluation on Duolingo's AI-generated lessons [13] from the perspective of its users on how it affects

their learning experience. With this understanding, we provide design considerations for language acquisition technologies that better cater to a student's individual professional circumstances. Specifically, we ask the following research questions (RQs):

RQ1: What are the differences in perceptions of language learning students towards general and work-specific scenarios from Duolingo's AI-generated lessons?, and

RQ2: How do these perceptions affect their learning experience towards gaining professional level fluency in their target language?

We conducted a formative study that surveyed five language learners employed by a multinational corporation about their experiences with Duolingo. They were asked questions about how often they encounter lessons containing scenarios that they directly experience in their work-related and general, non-work-related communications, and how these lessons affect their learning experience. They also provided specific lesson scenarios that would help them better gain professional-level communication fluency in their target language. We defined professional fluency as a level where *an individual can comfortably communicate various work-related topics in their target language.*

Our findings reveal that general scenarios provide a valuable learning experience for language learners, particularly for beginners. Their simplicity and relatability allow them to easily learn the foundational aspects of their target language, such as grammar and culture. However, work-related scenarios, especially ones that contain domain specific information, provide an opportunity for language learners to bridge the gap between the current fluency and professional-level fluency. Lastly, language learners expressed their desire for more personalization that caters to their learning goals and experience.

2 Background

2.1 Intelligent Tutoring Systems for Personalized Learning

Intelligent tutoring systems (ITSs) are computerized teaching tools that mimic human teaching behavior by using techniques from AI, cognitive science, and educational research. [12]. These ITSs are able to offer more personalized teaching methods because of the following modules that comprise its architecture: The *expert knowledge module* is the part of the system that contains domain-specific information and is responsible for generating lessons that are to be taught to the student [12, 14]. The *student model module* is responsible for diagnosing [17] the student's current understanding of the lesson material, and making necessary changes to the teaching medium to accommodate the student's needs [14]. Early work by Self [17] formalized these as *diagnostic* and *strategic* functions, respectively. Lastly, the *tutoring module* regulates the pedagogic interventions that will be presented to the student such as hints, tests, and explanations [12, 14]. ITSs can be very beneficial for a student. Prior work by Kulik and Fletcher [8] revealed that students who received tutoring from these ITSs outperformed students who didn't by a significant margin across different cultural and educational settings [8]. This is especially true more for local tests administered by specific instructional programs than standardized tests [7, 8, 16].

2.1.1 Duolingo as an Intelligent Tutoring System. An article by Parker [13] describes Duolingo's process of creating courses and lessons. A set of human experts writes a prompt with a set of corresponding rules that would be provided to the LLM to generate an exercise in the target language. These lessons are then evaluated by a human "learning designer" to see if they align with the language and the culture they represent. In the context of an ITS, this would be the expert knowledge module.

```
Write an exercise that uses the word VISITAR in SPANISH.

Rules:
1. The exercise must have two answer options.
2. The exercise must be fewer than 75 characters.
3. The exercise must be written in A2 CEFR level SPANISH.
4. The exercise must contain THE PRETERITE TENSE and THE IMPERFECT
TENSE.

Go!
```

Figure 2: A sample process of how Duolingo's "learning designers" prompts an LLM to generate exercises for its courses. (Image was grabbed from https://blog.duolingo.com/large-language-model-duolingo-lessons/)

Duolingo also offers some personalization for its users. It contains a student model [17] called "Birdbrain" that infers a student's expertise and adjusts the lesson's difficulty level accordingly [1], as well as provides certain lessons at specific phases of the learner's journey to maintain previously learned knowledge [18].

Lastly, Duolingo provides pedagogic interventions in the form of providing detailed feedback to the user when they answer lessons incorrectly [3]. This would be the tutoring module in an ITS.

2.2 Second Language Acquisition Theory

Second language acquisition is a field of study that aims to understand how individuals acquire a second language. A prominent theory in the field is one by Krashen [6]. His theory comprises of five hypotheses, but emphasizes the input hypothesis as the most important concept as it attempts to answer how people acquire language. It claims that people acquire a language when the input is *mostly* understood, meaning that there is some input that is beyond the current fluency level of the acquirer. The meaning of the input is not lost and is understood by the acquirer through context [6].In the case of Duolingo, the lessons it generates serves as the input, and it provides the context through the real-world scenarios that it presents. It also adjusts the difficulty for each lesson such that new lessons are slightly beyond the user's current fluency level [20].

In the realm of sociolinguistic theory, Dell Hymes developed the *Communicative Competence Theory*. Hymes et al. [4] posited that while grammatical knowledge of a language is important for an individual's acquisition, how an individual uses grammatical tools to construct sentences and participate in discourse is just as crucial. Hymes et al. [4] accounted for the differences in language use that occur because of both the context it is used in, and the varying background of its language users [19]. It is then, through the Communicative Competence Theory, that the measurement of a language learner's proficiency is not only based on their grammatical knowledge and syntactically accurate sentences, but also on how well they were able to communicate in different scenarios.

The work of Hymes et al. [4] led to a shift in the way second language acquisition was conducted. Initially, second language teachers emphasized grammatical structure and linguistic prescriptivism in their classrooms. This meant that the teacher's main goal was to pass on knowledge of the "right" use of the language. Following the propagation of Communicative Competence Theory, language teaching and learning eventually included examining what words, phrases, and structures were relevant to the context in which learners were speaking [9]. Social and hierarchical relationships, professional fields, and text types/genres (literary, academic, etc.) are just some examples of the different non-linguistic factors that can affect language use. By looking into how environment and context shape human interaction, the Communicative Competence Theory challenged what was deemed "appropriate" language use, by extending this definition to include both an utterances adherence to grammatical rules of a language and how well it communicated meaning based on social, professional, and even pragmatic context.

3 Method

We conducted a survey aiming to understand language learners' experiences with Duolingo. Duolingo was selected due to its popularity in the Philippines [15]. The survey was deployed on Qualtrics, and distributed through a multinational company based in the Philippines' communication channels. We received five valid responses. All participants were software engineers recruited from the company's Korean language class. The survey begins by asking participants about how long and how frequent they have used Duolingo for second language acquisition. Then for both general and work-related lesson scenarios provided by Duolingo, the survey asks how often they encounter them, and if they enhance or hamper their overall experience towards gaining professional-level fluency. Lastly, the participants suggested lesson scenarios that would enhance their learning experience.

4 Findings and Discussion

To better visualize our findings, we separate our design considerations according to our respondents' perceptions between lessons that contain *general scenarios* and *work-related scenarios*. We also describe some suggestions from our respondents on what type of lessons scenarios would help the easily gain professional-level fluency.

4.1 General scenarios serve their purpose for setting the foundation for the language acquisition

In the context of our study, general scenarios remain integral to participants' second language acquisition goals as all of them, especially self-described novices in the target language, report that these types of scenarios enhance their learning experience. Because these scenarios are relatable, and are encountered more frequently in their daily lives, it enabled them to grasp foundational language concepts such as grammar and vocabulary easily. This is shown in some responses to Q7: *"The non-work-related lessons help in adding to my overall learning, specifically with grammar and vocabulary."* and *"I can identify the words I hear during daily conversations"*. Duolingo currently satisfies this, as all respondents reported encountering these general scenarios more often than work-specific scenarios.

4.2 Work-related scenarios serve as an opportunity to bridge the fluency gap between novice and professional

We find that respondents encounter work-related scenarios much less frequently than general ones. One respondent even reported having never encountered any, but for those who did, the general perception is that the work-specific scenarios are able to bridge the fluency gap between novice and professional. One respondent reported that learning about work-specific scenarios helped him/her better understand culture and language. In a more specific instance, one respondent, who we presume is employed as a software developer, reported that he/she has yet to encounter scenarios that contain technical jargon such as CI/CD pipeline, user interface, and defect, widening his/her perceived learning gap towards being able to communicate in professional settings.

However, we propose that these work-related scenarios may be more beneficial in later stages of their learning journey. One respondent noted that work-related scenarios are still not applicable at the beginner level, showing that students are aware that it is necessary to understand the fundamentals of the language from general scenarios before they can acquire professional-level fluency from work-related scenarios.

4.3 Divergent topics on suggested lesson scenarios show a desire for a more personalized learning experience

In our survey, we asked for suggestions on what type of lessons would help them gain professional level fluency. The suggestions' topics were divergent, such as: more language fundamentals, scenarios on negotiating task deadlines, traveling to the target language's country, and more everyday conversations. These topics suggest that there is an opportunity to leverage participants' backgrounds, goals, and habits to design language learning applications that can create more personalized lesson scenarios. Because Duolingo's Birdbrain/student modules only adjusts for the difficulty level, it does not adjust according to the lesson content that the individual learner desires.

5 Limitations and Future Work

Our work contains limitations. We recognize the small pool of participants, and plan to continue this study by recruiting more participants to strengthen our findings. For our future experiment, we plan to fine-tune an LLM to generate lessons that are more applicable in the technology industry. We will then conduct a long-term between-subjects study with software engineers as language learners. Our control group would be one without the ITS, the other group would use Duolingo, and the last one would use our fine-tuned LLMs. We will then compare language test scores and reported user experience between the groups.

6 Conclusion

We present a formative study evaluating Duolingo's LLM-generated lessons from the language learners' perspective. We found that learners perceive lessons with general scenarios as relatable and, therefore, helpful for them as novices to learn more foundational concepts of the language through immersion. However, the benefits of these general scenarios may not transfer well for their professional career settings, where scenarios typically involve very domain-specific technical conversations. It would be beneficial for language learning applications to provide both general scenarios and adapt to these types of individual work-specific scenarios to provide a more personalized learning experience in general. In closing, we envision an intelligent language tutoring system that is agentic. One that is able to understand and adapt to the user's changing background, goals, context, and environment, to create lesson content that caters to the each unique individual learner.

7 Appendices

7.1 Survey Questions

Q1 (Single Choice) How long have you been using Duolingo for second language acquisition?
- < 1 year
- 1-5 years
- 6-10 years
- > 10 years

Q2 (Single Choice) How frequently do you use Duolingo for second language acquisition?
- Less than once a month
- Once a month
- 2-3 times a month
- Once a week
- 2-3 times a week
- Daily

Info: In language learning settings, real-world scenarios in the target language are commonly presented to the learners, providing a level of immersion to help them gain professional-level fluency in the target language.
(we define professional-level fluency as where an individual who is able to comfortably communicate in the target language various work-related matters)

Q3 (Single Choice) How frequently do you encounter lessons containing scenarios that you directly experience in your work-related communications?
- Always
- Very Often
- About half the time
- Sometimes
- Never

Q4 (Single Choice) Do these work-related lessons enhance/hamper your overall experience towards gaining professional-level fluency in your target language?
- Greatly enhances
- Somewhat enhances
- Neutral
- Somewhat hampers
- Greatly hampers

Q5 (Open Ended) Please provide a brief explanation on why you responded such in the previous question

Q6 (Single Choice) How frequently do you encounter lessons containing scenarios that you experience in your non-work-related communications?
- Always
- Very Often
- About half the time
- Sometimes
- Never

Q7 (Single Choice) Do these non-work-related lessons enhance/hamper your overall experience towards gaining professional-level fluency in your target language?
- Greatly enhances
- Somewhat enhances
- Neutral
- Somewhat hampers
- Greatly hampers

Q8 (Open Ended) Please provide a brief explanation on why you responded such in the previous question

Q9 (Open Ended) What type of lesson scenario/s would help you easily gain professional level fluency in your target language?

8 Appendices

8.1 Survey Results

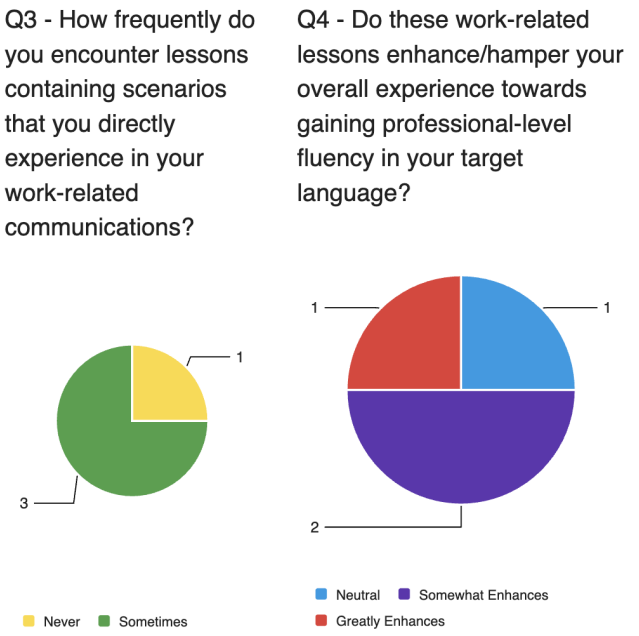

Figure 3: Survey results of questions on language learners' perceptions on lessons containing general scenarios)

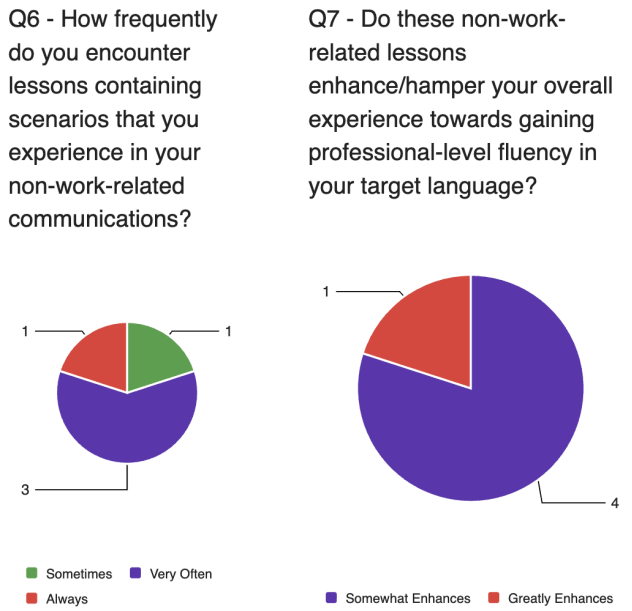

Figure 4: Survey results of questions on language learners' perceptions on lessons containing work-specific scenarios)

References

[1] Klinton Bicknell and Claire Brust. 2020. Learning how to help you learn: Introducing Birdbrain! https://blog.duolingo.com/learning-how-to-help-you-learn-introducing-birdbrain/

[2] Cindy Blanco. 2025. 2025 Duolingo Language Report. https://blog.duolingo.com/2025-duolingo-language-report/

[3] Luis Castillo. 2026. Explain My Answer is now free for all learners! https://blog.duolingo.com/explain-my-answer-now-free/

[4] Dell Hymes et al. 1972. On communicative competence. *sociolinguistics* 269293 (1972), 269–293.

[5] Stephen Krashen. 1981. Second language acquisition. *Second Language Learning* 3, 7 (1981), 19–39.

[6] Stephen Krashen. 1982. Principles and Practice in Second Language Acquisition. (1982).

[7] Chen-Lin C. Kulik, James A. Kulik, and Robert L. Bangert-Drowns. 1990. Effectiveness of Mastery Learning Programs: A Meta-Analysis. *Review of Educational Research* 60, 2 (1990), 265–299. arXiv:https://doi.org/10.3102/00346543060002265 doi:10.3102/00346543060002265

[8] James A. Kulik and J. D. Fletcher. 2016. Effectiveness of Intelligent Tutoring Systems: A Meta-Analytic Review. *Review of Educational Research* 86, 1 (2016), 42–78. arXiv:https://doi.org/10.3102/0034654315581420 doi:10.3102/0034654315581420

[9] T.M. Lillis. 2006. Communicative Competence. In *Encyclopedia of Language Linguistics (Second Edition)* (second edition ed.), Keith Brown (Ed.). Elsevier, Oxford, 666–673. doi:10.1016/B0-08-044854-2/01275-X

[10] Reza Hadi Mogavi, Chao Deng, Justin Juho Kim, Pengyuan Zhou, Young D Kwon, Ahmed Hosny Saleh Metwally, Ahmed Tlili, Simone Bassanelli, Antonio Bucchiarone, Sujit Gujar, et al. 2024. ChatGPT in education: A blessing or a curse? A qualitative study exploring early adopters' utilization and perceptions. *Computers in Human Behavior: Artificial Humans* 2, 1 (2024), 100027.

[11] Ethan R Mollick and Lilach Mollick. 2023. Using AI to implement effective teaching strategies in classrooms: Five strategies, including prompts. *The Wharton School Research Paper* (2023).

[12] Hyacinth S Nwana. 1990. Intelligent tutoring systems: an overview. *Artificial Intelligence Review* 4, 4 (1990), 251–277.

[13] Henry Parker. 2023. How Duolingo uses AI to create lessons faster. https://blog.duolingo.com/large-language-model-duolingo-lessons/

[14] Martha C Polson and J Jeffrey Richardson. 2013. *Foundations of intelligent tutoring systems.* Psychology Press.

[15] Ralph Rivas. 2022. WATCH: Filipinos use Duolingo more during pandemic for K-drama, anime binge. *Rappler* (2022). https://www.rappler.com/business/video-filipinos-use-duolingo-covid-19-pandemic-korean-drama-anime-binge/

[16] Barak Rosenshine and Carla Meister. 1994. Reciprocal Teaching: A Review of the Research. *Review of Educational Research* 64, 4 (1994), 479–530. arXiv:https://doi.org/10.3102/00346543064004479 doi:10.3102/00346543064004479

[17] John Self. 1988. Student models: what use are they. *Artificial Intelligence Tools in Education* (1988), 73–86.

[18] Burr Settles and Brendan Meeder. 2016. A trainable spaced repetition model for language learning. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers).* 1848–1858.

[19] Shona Whyte. 2019. Revisiting communicative competence in the teaching and assessment of language for specific purposes. *Language Education & Assessment* 2, 1 (2019), 1–19.

[20] Sharon Wilkinson. 2024. Dear Duolingo: What's the right level of difficulty? https://blog.duolingo.com/right-level-of-difficulty/

[21] Ying Zheng, Shuyan Huang, Xiaoli Zeng, Yaying Huang, Zitao Liu, and Weiqi Luo. 2025. Knowledge-enhanced large language models for automatic lesson plan generation. *Humanities and Social Sciences Communications* 12, 1 (2025), 1784.