LLM chatbots as a language practice tool: a user study

Gladys Tyen, Andrew Caines, Paula Buttery
ALTA Institute, Dept. of Computer Science & Technology
University of Cambridge
{gladys.tyen, andrew.caines, paula.buttery}@cl.cam.ac.uk

Abstract

Second language learners often experience language anxiety when speaking with others in their target language. As the generative capabilities of Large Language Models (LLMs) continue to improve, we investigate the possibility of using an LLM as a conversation practice tool. We conduct a user study with 160 English language learners, where an LLM chatbot is used to simulate real-world conversations. We present our findings on 1) how an interactive session with a chatbot might impact performance in real-world conversations; 2) whether the learning experience differs for learners of different proficiency levels; 3) how changes in difficulty affects the learner's experience; and 4) how online, synchronous conversation provided by an LLM compares with a purely receptive experience. Additionally, we propose a simple yet effective way to detect linguistic complexity on-the-fly: clicking on words to reveal dictionary definitions. We demonstrate that clicks correlate well with linguistic complexity and indicate which words learners find difficult to understand.

1 Introduction

Rapid advancements in natural language processing technology, brought on by large language models (LLMs), have opened up new directions and methods for learning and education. In particular, language learners have been making use of LLMs' language generation abilities to support their learning experience (e.g. PrettyPolly, 2023; Microsoft, 2023).

In this paper, we investigate the possibility of using an LLM for conversational practice in language learning. Many existing approaches restrict the LLM in some way (e.g. Duolingo Team, 2023; Zhang and Huang, 2024), requiring manual crafting of prompts or syllabuses. Restrictions are common for pre-LLM chatbots in language learning (Bibauw et al., 2019), as they are rule-based and can often fail to parse user input correctly. However, as LLM technology advances, these restrictions may no longer be needed.

In our study, we test the limits of LLM capabilities by using an LLM directly without any restrictions on topic, context, or grammatical form. We conduct a user study with 160 English learners, who are asked to interact with an online chatbot. In our implementation, our chatbot is designed to simulate a typical conversationalist so that the learner can practise chatting in English.

We seek to answer the following research questions:

RQ1. Does chatting with an online chatbot have any educational impact on real-life interaction?

RQ2. How does the language learning experience change for learners at different proficiency levels?

RQ3. Does adjustment of difficulty level affect the learner's experience, either positively or negatively?

RQ4. How does a conversational setting (combining comprehension and production) compare to a comprehension-only setting?

Overall, our results suggest that chatbots for conversational practice have positive educational impact, though further investigation is required in some areas. We find that this setup is more suited to learners at lower proficiency levels; that it provides more enjoyment over plain reading; and that personalised difficulty adaptation prevents dialogues from becoming too easy. Detailed findings can be found in Section 4.

Additionally, we propose a simple but effective way to identify linguistic complexity during

Gladys Tyen, Andrew Caines and Paula Buttery. LLM chatbots as a language practice tool: a user study. *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2024)*. Linköping Electronic Conference Proceedings 211: 235–247.

a chatbot conversation: clicking to reveal dictionary definitions. This function can be seamlessly integrated into any web interface, and our results demonstrate a clear correlation between clicking and what the learner finds difficult.

2 Background

Before the advent of transformers and LLMs, chatbots for computer-assisted language learning (CALL) were typically rule-based and were only used for constrained scenarios. Bibauw et al. (2019) present a pre-LLM survey of dialogue systems for language learning, and observe that most systems have implicit or explicit constraints, on either the content of the user response or the grammatical form. Ones that allow free dialogue are typically rule based and prone to producing ungrammatical or nonsensical messages (e.g. Coniam, 2014; Jia, 2009)

However, as most chatbots worked within these constraints, it was also easier to introduce adjustments to the chatbot for language learning purposes. One of the most common adjustments is the adaptation of difficulty level based on the user's linguistic proficiency or previous performance, for example as implemented by Hassani et al. (2016); Lu et al. (2006); Ní Chiaráin and Ní Chasaide (2016); Su et al. (2015); Vlugter et al. (2009).

With the introduction of neural dialogue systems and later LLMs, the performance of chatbots improved greatly (Papangelis et al., 2021; Adiwardana et al., 2020; Roller et al., 2021). This technology made it possible to build chatbots for CALL with little to no constraints, while generating grammatical sentences. For example, Tyen et al. (2022) propose a chatbot setup where the difficulty of generated text can be adjusted to user's proficiency level; Lee et al. (2023) propose a system (with some restriction on context) that produces feedback for students; Zhang and Huang (2024) investigate how vocabulary acquisition is affected by 4 types of chatbots for 4 contexts, all connected to an LLM backend. Additionally, the release of ChatGPT (OpenAI, 2023) prompted some language learners to use the service to help them learn (Microsoft, 2023), even though ChatGPT is not specifically designed for language learning.

Despite advances in technology and commercial chatbots for language learning, there is limited research on the effect of using unconstrained LLM chatbots to learn a second language. Previous studies use chatbots that are limited to predetermined contexts (Lee et al., 2023; Zhang and Huang, 2024), or that are rule-based (Coniam, 2014; Jia, 2009), with the feedback that the chatbot is difficult to understand or responds with ungrammatical or nonsensical messages.

In our paper, we use an open-domain LLM chatbot, with no restrictions on context, topic, or grammatical form. Our chatbot is designed to simulate a typical conversationalist, so that learners may practise conversing in their target language. To our knowledge, this work is the first to perform user evaluations on open-domain LLM chatbots for language learning.

3 Study setup

We recruit 160 participants via Prolific[1] for our user study. All participants are screened to ensure that their first language is not English. They are then directed to our website, where they navigate through 4 sections:

1. The first section consists of basic profiling questions to ascertain the participant's linguistic background, such as their first language (L1). The most common L1s were Polish, Portuguese, and Italian (full list in the appendix).

2. The second section is a proficiency test consisting of 25 multiple choice questions to estimate their proficiency level. The questions and answers are taken from the Cambridge English Test Your English application[2]. Scores from the test are mapped to the Common European Framework of Reference (CEFR) (Council of Europe, 2020), a 6-point scale representing proficiency, allowing easy comparison with existing work.

3. The third section is the main interaction with the chatbot. This involves chatting directly with the chatbot, or reading messages from chatbots; variations are described below.

4. The final section consists of closing questions asking the participant about their experience, including 2 attention questions to eliminate low-effort responses. We enclose the full

[1] https://www.prolific.com/

[2] https://www.cambridgeenglish.org/test-your-english/general-english/

list in the appendix, but highlight individual questions in our Findings section.

Additional details of the user study setup can be found in the appendix.

Each participant is randomly assigned different experimental conditions in a $2 \times 2 \times 2$ design:

- **Chatting VS reading**
  To understand the difference between receptive reading and interactive conversation, we assign half of our participants to the chatting condition, and the remaining half to the reading condition. In the chatting condition, each participant is asked to converse with a chatbot. They send messages to the chatbot directly and can actively steer the conversation topic. In the reading condition, the participant cannot send messages, and instead navigates through a conversation between two identical chatbots. Everything else, such as the user interface, remains the same.

- **Adaptive difficulty VS non-adaptive difficulty**
  One common feature in language learning chatbots is the capability of adapting chatbot messages based on the user's proficiency level. However, it is unclear to us how this may affect the learning experience, so we apply the adaptation for half of the participants, while the other half receive messages generated with standard top-k sampling ($k = 40$) (Fan et al., 2018). For the adaptation, we follow Tyen et al. (2022) and use a reranking method with sub-token penalties and filtering, as described in their paper[3]. See the appendix for further details on the re-ranking model and implementation of penalties.

- **Dictionary lookup VS no dictionary lookup**
  In the dictionary lookup condition, participants are able to click on words to look up their definitions. This function is only available for words in messages that are sent from the chatbot. All messages are tokenised by the RASP parser (Briscoe et al., 2006).

Full details can be found in the appendix.

[3]Implementation found at https://github.com/WHGTyen/ControllableComplexityChatbot.

For all three pairs of conditions, participants are split evenly into two groups, where one group is assigned one condition and the other group is assigned the other condition: for example, there are 80 participants in the chatting condition and 80 participants in the reading condition as well. The splitting is done in a way that ensures equal coverage across all combinations of conditions: e.g. there are 20 participants who are chatting *and* have adaptive difficulty *and* dictionary lookup; 20 participants who are reading *and* have adaptive difficulty *and* dictionary lookup; and so on.

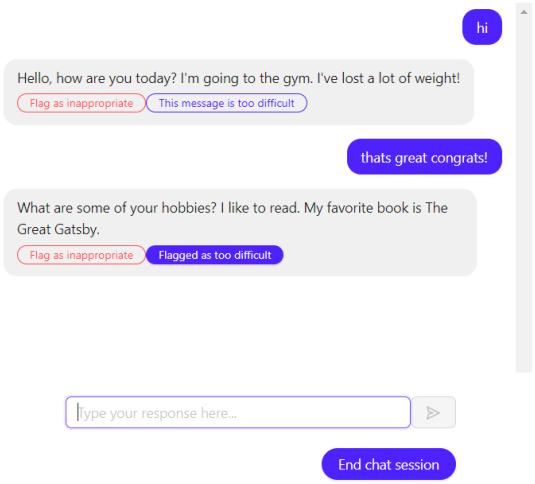

Figure 1: Chat interface presented to participants in the chatting condition. Messages in blue bubbles are sent from the user, while messages in grey bubbles are sent from the chatbot. In this example, the most recent message is flagged by the user as being too difficult.

3.1 Chatbot

We use BlenderBot (2.7B parameters) (Roller et al., 2021) as the base LLM. BlenderBot was chosen because the model is not instruction-tuned, and has been fine-tuned on the Blended Skill Talk dataset (Smith et al., 2020), which combines various conversational skills. This allows us to simulate a real conversationalist rather than a virtual assistant. Additionally, BlenderBot was previously used by Tyen et al. (2022) for difficulty adjustment. We use the same setup[3] to enable a clear comparison: for participants in the adaptive condition, we use a decoding method proposed by Tyen et al. (2022) (method 5), which allows us to adjust the difficulty level of generated messages.

In terms of chatbot quality, Roller et al. (2021) report extensive evaluation results on BlenderBot, including self-chat human evaluation and interac-

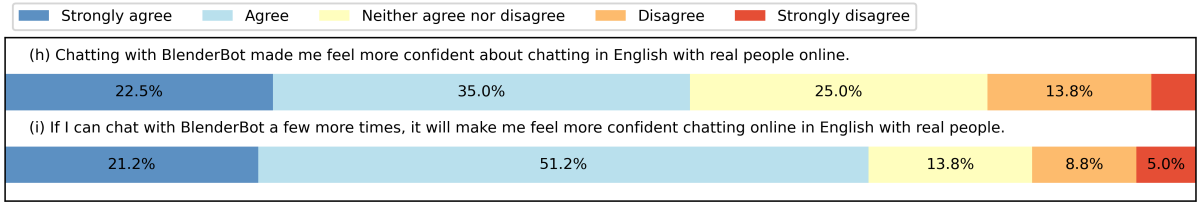

Figure 2: Responses to confidence-related Likert questions from participants in the chatting condition.

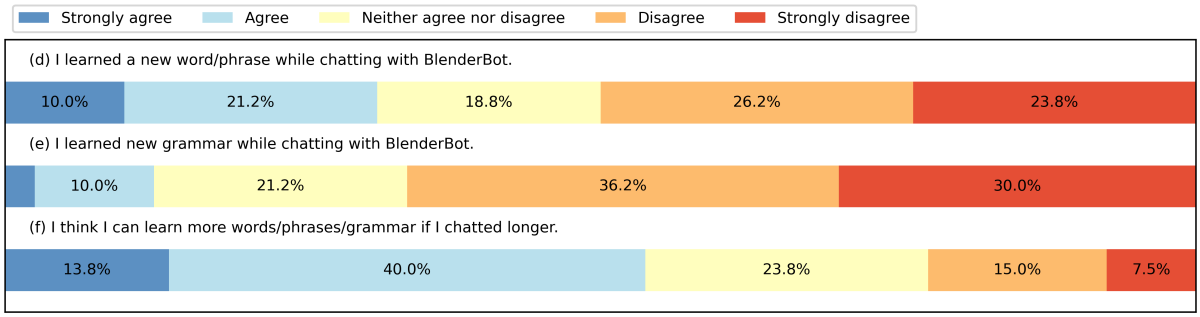

Figure 3: Responses to learning-related Likert questions from participants in the chatting condition.

tive human evaluation. Their results show that generative BlenderBot (2.7B) performs better than Meena Adiwardana et al. (2020) and narrowly loses to human participants in terms of engagingness (49% versus 51%). Tyen et al. (2022) report self-chat evaluation results of the adapted decoding method based on the Sensibleness and Specificity Average Adiwardana et al. (2020) and grammaticality. Method 5 from their paper was found to be statistically equivalent to the non-adapted version in terms of sensibleness, specificity, as well as grammaticality.

Additionally, to disentangle effects of prompt crafting or manual changes to the learning experience, and to minimise effects on chatbot quality, our current chatbot setup does not use any prompts, predetermined responses, or linguistic syllabuses (though they may be added in future work). All user input goes directly to the LLM, and all generated messages are sent directly to the user.

Figure 1 shows a screenshot of the interface used to interact with the chatbot. Participants are asked to spend at least 15 minutes on this section, after which the "End chat session" button would appear. Participants can also choose to spend more time with the chatbot if they wished.

4 Findings

4.1 RQ1: Impact on real-life interaction

Increased self-confidence in real-life interaction Two of our feedback questions (h) and (i), shown in Figure 2, focus on the learner's sense of self-confidence when it comes to real-life settings. We rely on self-reports as confidence is inherently about perception of the self, and arguably can only be measured via self-reports (Paulhus et al., 2007).

The results show that more than half of the participants in the chatting condition agree that they felt more confident about chatting with real people, even after 1 session of conversing with the chatbot. This number increases further to 72% in question (i), where we ask participants for *predicted* self-confidence levels, if given more opportunities to converse with the chatbot.

Limited learning may increase in the long term Questions (d), (e), and (f), shown in Figure 3, focus on the learning of new words, phrases, or grammatical constructions. While some participants report learning after just one session, most disagree with the statements, particularly regarding grammatical constructions. This suggests that a single chatbot session is unlikely to provide benefits for language learning.

On the other hand, participants are more optimistic when asked to *predict* learning, if given fur-

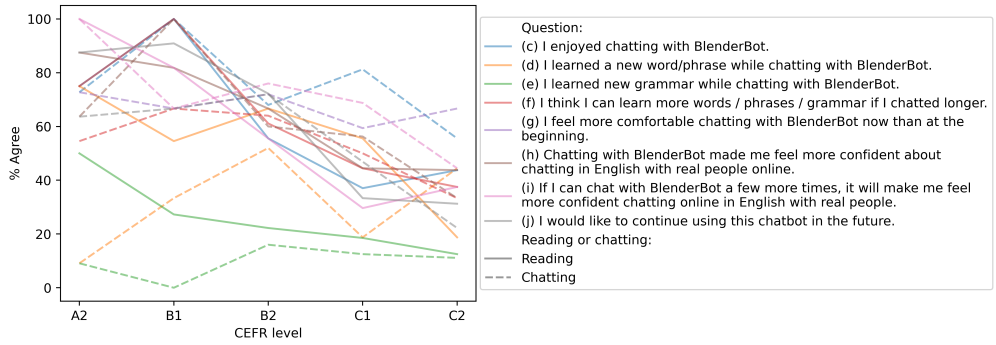

Figure 4: Proportion of *Agree* or *Strongly agree* responses to each Likert question, sorted by CEFR level.

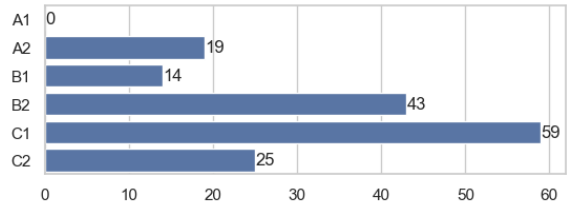

Figure 5: Distribution of CEFR levels across all participants. CEFR levels are ordered from least to most proficient.

ther opportunity to converse with the chatbot. This is in line with our previous finding about confidence, where participants also predict more positive outcomes if given more time with the chatbot. As our user study only consists of one session and is not designed to test longitudinal effects, we are unable to verify whether there are any actual long term benefits. However, it is noteworthy that users themselves have a positive opinion on long-term chatbot usage, suggesting that their experience had a motivational effect.

4.2 RQ2: Variation in proficiency levels

All participants are asked to complete a series of multiple choice questions, which are used to gauge their proficiency level. The distribution of CEFR levels is shown in Figure 5. None of our participants are found to be at A1 (most beginner) level: this is likely due to the initial recruitment and navigation through the consent form, which requires a minimal level of proficiency to understand.

Proportion of *Agree* or *Strongly agree* responses sorted by approximate CEFR level are visualised in Figure 4. Note that *Agree* and *Strongly agree* represent positive outcomes in our Likert questions, while *Disagree* and *Strongly disagree* represent negative outcomes.

We then compute Spearman's rank correlation coefficient (ρ) between test scores and answers to our Likert questions.

Our results show that **less proficient learners are more likely to report and predict positive outcomes**. We find that participants' scores in the proficiency test significantly negatively correlate with:

- enjoyment (question (c), $\rho = -0.25$, $p < 0.002$)
- perceived learning of grammatical constructions (question (e), $\rho = -0.33$, $p < 0.00003$)
- predicted learning in the long term (question (f), $\rho = -0.27$, $p < 0.0005$)
- predicted self-confidence levels in the long term (question (i), $\rho = -0.36$, $p \ll 0.00001$)
- interest in continued usage (question (j), $\rho = -0.37$, $p \ll 0.00001$)

Questions (f), (i), and (j) all pertain to participants' predictions, suggesting that lower proficiency participants find greater potential for future benefits than high-proficiency participants. This is a reasonable outcome as more beginner language learners would require more practice than more experienced learners. For question (e), we hypothesise that the difference between high- and low-proficiency learners is because grammar is often taught at earlier stages of learning. High-proficiency learners are more likely to struggle with advanced concepts such as use of humour and slang, linguistic style, etc.

Note that the above correlation scores are computed for all participants (in the reading and chatting conditions). Figure 4 shows that the effect is stronger for the reading condition than the chatting condition, where learners at a higher proficiency level give more positive responses than in the reading condition, particularly for questions (c) on enjoyment and (i) on predicted confidence.

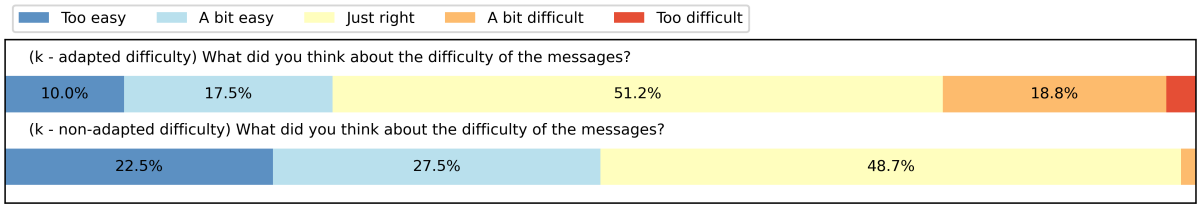

Figure 6: Responses to question (k) on perceived difficulty from participants in the adapted and non-adapted difficulty conditions.

Overall, our results suggest that learners at a lower proficiency level are more likely to benefit from interactions with an LLM chatbot, but correlations are not strong and many high-proficiency learners also report positive outcomes.

4.3 RQ3: Difficulty adaptation

For half of the participants, the chatbots are adjusted to their CEFR level (Tyen et al., 2022) based on their scores in the pre-test. For the remaining half, the chatbots use standard top-k sampling (Fan et al., 2018). At the end of the study, participants are asked about the difficulty level of the messages in question (k), where the potential responses are: *Too easy*, *A bit easy*, *Just right*, *A bit difficult*, and *Too difficult*.

Firstly, our results show that there is a significant difference in perceived difficulty between those in the adapted condition and those in the non-adapted condition ($p < 0.00009$). When comparing specific responses, we find that participants in the adapted version are **significantly more likely to respond with *A bit difficult*** ($p < 0.0001$), while the number of responses for *Too difficult* remain the same, and there are non-significant reductions in the number of *Too easy* and *Easy* responses. Figure 6 contains a visualisation of the responses.

The fact that the non-adapted version of the chatbot is *Too easy* for many participants is in line with the finding in Tyen et al. (2022) that BlenderBot with no adaptations generates messages at B1 level. If the default difficulty level is B1, many participants at B2 level or above would consider the messages to be too easy. Therefore, difficulty adjustment methods are required.

Our results indicate that difficulty adjustment via decoding (Tyen et al., 2022) is effective at introducing language aspects which are more difficult, but are not so difficult that the learner is unable to comprehend it. According to Krashen's Input Hypothesis of second language acquisition (Krashen, 1992), successful second language acquisition occurs when the learner is exposed to input that contains '$i + 1$', referring to "an aspect of language that the acquirer has not yet acquired but that he or she is ready to acquire". This suggests that the ideal perceived difficulty level is between *Just right* and *A bit difficult*. Following this hypothesis, we surmise that exposure to text with adjusted difficulty levels is likely more beneficial for second language learning than to text that is not adjusted. However, to fully test this theory, a longitudinal study is required to measure learning progress.

4.4 RQ4: Conversational interaction versus receptive reading

In both the chatting condition and reading condition, messages from the chatbot(s) are generated on-the-fly using the same decoding strategy. Despite using the same setup, we observe distinct linguistic differences between the content generated in the chatting and reading conditions, likely due to influence from the user. For example, messages generated in the reading condition are shorter on average ($p < 0.0002$); messages in the chatting condition are more likely to contain questions ($p << 0.00001$).

Overall, the Jaccard similarity between chatbot-generated messages in the chatting and reading conditions is relatively high at 0.35. For comparison, the Jaccard similarity between all chatbot-generated messages and messages in the Blended Skill Talk dataset (Smith et al., 2020) (which BlenderBot was fine-tuned on) is 0.26; and the Jaccard similarity between chatbot-generated messages and user-written messages in the user study is 0.12.

We additionally explore the impact of reading versus chatting via survey responses. Surprisingly, our results show only one main difference between learners in the chatting and reading conditions: **chatters enjoy the experience more than read-**

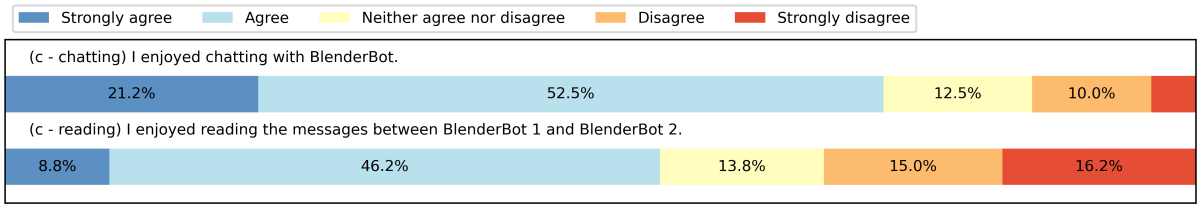

Figure 7: Responses to Likert question on enjoyment from participants in the chatting and reading conditions.

ers do. Figure 7 below shows a comparison of their responses to question (c), which asks whether participants enjoyed the chatbot session. Chatters are significantly more likely to give a more positive response ($p < 0.001$).

Among all survey question responses, other than enjoyment, we find no other significant differences between the chatting and reading conditions, whether with adaptive or non-adaptive difficulty, or with or without dictionary lookup. This is a surprising result given the differences in text content, and the fact that second language production is inherently differently from second language comprehension (Laufer, 1998; Gernsbacher and Kaschak, 2003).

There are some suggestive, but non-significant differences: for example, users in the chatting condition are slightly more likely to predict boosts in confidence levels, while users in the reading condition are slightly more likely to report learning new words. However, further study with a larger group of users is required to understand if these effects are linked to interaction (or lack thereof).

5 Clicking for dictionary lookup as an indicator of complexity

In our user study, we implement a clicking mechanism where learners can click on words to reveal their dictionary definition. This function is simple to implement and integrates seamlessly with the existing user interface, yet can provide valuable information about the user's learning experience.

We find that clicks are a strong indicator of when a learner finds a word difficult. We report in Table 1 three statistics that are often correlated with lexical complexity (Shardlow et al., 2021), and compare them for words that are clicked on versus words that are *not* clicked on. We find that words that are clicked on are more complex, as they are significantly longer ($p << 0.0001$), less frequent ($p << 0.0001$), and have a smaller number of definitions ($p < 0.0002$).

Statistic	Clicked	Unclicked
Avg. character length	**8.07**	3.80
Avg. Zipf frequency	**5.69**	6.82
Avg. num. of definitions	**2.59**	5.26

Table 1: Statistics correlated with lexical complexity for words that are clicked on, versus words that are not clicked on. Zipf frequency refers to the base-10 logarithm of frequency per 1 billion words; the number of definitions refers to the number of synsets on WordNet (Miller, 1994). Bold font denotes the statistic that indicates higher complexity. All 3 statistics are shown to be significantly different between clicked and unclicked words ($p << 0.0001$ for length and frequency; $p < 0.0002$ for number of definitions).

Furthermore, clicks are also associated with the reported difficulty level of the overall message. During our study, participants are able to flag messages that they consider to be too difficult (see Figure 1). We find that messages that are flagged as difficult are 5 times more likely to have words that are clicked on (11.3%), compared to messages that are not flagged (2.2%). This demonstrates that learners are clicking on words that *they* consider complex, rather than e.g. out of curiosity, or due to random, unintentional clicking.

Despite strong evidence that clicks are indicative of lexical complexity, we observe that only 33 out of 80 participants in the clicking condition make use of this feature. For the 33 participants, 4751 messages are sent from the chatbot, but only 377 clicks are recorded in total. Possible reasons for the low click-rate include: 1) Participants rarely encounter any words that they find sufficiently difficult; 2) Participants are engaged in conversation and prefer to continue rather than pausing to read definitions; 3) Participants find the dictionary definitions unhelpful; or 4) Participants forget they have access to this function. Note that all participants in the clicking condition are informed of this mechanism before their chatbot session.

Due to the low click-rate, our data is insufficient to draw conclusions about potential benefits

or drawbacks of clicking. Additionally, we find no significant differences in survey response questions between the groups with and without this dictionary lookup function. This is also the case when looking at groups with or without adaptive difficulty, or in the chatting or reading conditions. Further work is required to understand clicking behaviour and its impact on the learning experience.

6 Limitations and future work

Scope of user study Our user study involves a small sample of 160 participants, whose first languages are mostly European languages, and whose CEFR proficiency levels are skewed towards the higher end. Additionally, due to the small number of participants, we are unable to properly measure interaction effects despite the $2\times2\times2$ design. Further work is required to ascertain if our findings hold at a larger scale and with a different population, and to clarify how LLM chatbots facilitate language learning.

Measured performance Some of our observations rely on participants' self reports rather than measured linguistic performance. Based on previous research, our results show promise and are likely associated with improved performance, but our study does not measure this directly. In future work, we can measure linguistic improvement over the course of multiple chatbot sessions by comparing performance before and after the fact.

LLM capability For our user study, we use a small (2.7B parameters) model for the ease of deployment and inference speed. It is possible to improve the capability of the chatbot by replacing it with larger models such as LLaMA (Touvron et al., 2023) and BLOOM (BigScience Workshop et al., 2022). We expect that results related to enjoyment are likely to improve with a larger model, and the conversational experience would be more realistic.

Personalisation using clicking data Our current study does not make use of the clicking data to adjust the generated messages, but future work on computer-assisted language learning can make use of clicks to adapt content on-the-fly to the user.

7 Conclusion

In this paper, we report our findings from our user study, where we recruit 160 second language speakers to interact with LLM-based chatbots. Our results show that using an LLM chatbot as a language practice tool can improve self-confidence, and provides a more enjoyable learning experience compared to purely receptive reading tasks. Although learning outcomes are not apparent after one session, many participants predict more positive effects in the long term, if given further opportunity to interact with the chatbot. This is especially true for learners at a lower proficiency level.

In terms of implementation, we introduce clicking as a way to reveal dictionary definitions during the user study. We find that this method effectively detects words which the learner finds complex, on-the-fly. For the chatbot, we implement a decoding method that adjusts the difficulty of generated messages (Tyen et al., 2022). Our results show that this method generates text that is more often considered *A bit difficult*, which is likely to facilitate learning (Krashen, 1992).

Overall, our findings demonstrate that LLM chatbots as a language practice tool can bring benefits to different aspects of language learning. We leave it to future work to measure long-term learning outcomes of chatbot interaction.

Acknowledgements

This paper reports on research supported by Cambridge University Press & Assessment. We thank the anonymous reviewers for their comments.

References

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Serge Bibauw, Thomas François, and Piet Desmet. 2019. Discussing with a computer to practice a foreign language: research synthesis and conceptual framework of dialogue-based CALL. *Computer Assisted Language Learning*, 32(8):827–877.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel

Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Ted Briscoe, John Carroll, and Rebecca Watson. 2006. The second release of the RASP system. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 77–80, Sydney, Australia. Association for Computational Linguistics.

David Coniam. 2014. The linguistic accuracy of chatbots: usability from an esl perspective. *Text & Talk*, 34(5):545–567.

Council of Europe. 2020. *Common European Framework of Reference for Languages: Learning, Teaching Assessment*, 3rd edition. StrasBourg.

Duolingo Team. 2023. Introducing duolingo max, a learning experience powered by gpt-4. Accessed on January 31, 2024.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Morton Ann Gernsbacher and Michael P. Kaschak. 2003. Neuroimaging studies of language production and comprehension. *Annual Review of Psychology*, 54(1):91–114.

Kaveh Hassani, Ali Nahvi, and Ali Ahmadi. 2016. Design and implementation of an intelligent virtual environment for improving speaking and listening skills. *Interactive Learning Environments*, 24(1):252–271.

Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, et al. 2023. Ai alignment: A comprehensive survey. *arXiv preprint arXiv:2310.19852*.

Jiyou Jia. 2009. An ai framework to teach english as a foreign language: Csiec. *Ai Magazine*, 30(2):59–59.

Stephen Krashen. 1992. The input hypothesis: An update. *Linguistics and language pedagogy: The state of the art*, pages 409–431.

Batia Laufer. 1998. The Development of Passive and Active Vocabulary in a Second Language: Same or Different? *Applied Linguistics*, 19(2):255–271.

Seungjun Lee, Yoonna Jang, Chanjun Park, Jungseob Lee, Jaehyung Seo, Hyeonseok Moon, Sugyeong Eo, Seounghoon Lee, Bernardo Yahya, and Heuiseok Lim. 2023. PEEP-talk: A situational dialogue-based chatbot for English education. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 190–207, Toronto, Canada. Association for Computational Linguistics.

Chun-Hung Lu, Guey-Fa Chiou, Min-Yuh Day, Chorng-Shyong Ong, and Wen-Lian Hsu. 2006. Using instant messaging to provide an intelligent learning environment. In *Intelligent Tutoring Systems*, pages 575–583, Berlin, Heidelberg. Springer Berlin Heidelberg.

Microsoft. 2023. How ChatGPT can be used to help with foreign language learning. Accessed on February 1, 2024.

George A. Miller. 1994. WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Neasa Ní Chiaráin and Ailbhe Ní Chasaide. 2016. The digichaint interactive game as a virtual learning environment for irish. In *CALL communities and culture – short papers from EUROCALL 2016*, pages 330–336. Research-publishing.net.

OpenAI. 2023. ChatGPT. Accessed on February 1, 2024.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Alexandros Papangelis, Paweł Budzianowski, Bing Liu, Elnaz Nouri, Abhinav Rastogi, and Yun-Nung Chen, editors. 2021. *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*. Association for Computational Linguistics, Online.

Delroy L Paulhus, Simine Vazire, et al. 2007. The self-report method. *Handbook of research methods in personality psychology*, 1(2007):224–239.

PrettyPolly. 2023. Prettypolly - learn a language by practicing speaking with ai. Accessed on January 31, 2024.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.

Matthew Shardlow, Richard Evans, Gustavo Henrique Paetzold, and Marcos Zampieri. 2021. SemEval-2021 task 1: Lexical complexity prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1–16, Online. Association for Computational Linguistics.

Erfan Shayegani, Md Abdullah Al Mamun, Yu Fu, Pedram Zaree, Yue Dong, and Nael Abu-Ghazaleh. 2023. Survey of vulnerabilities in large language models revealed by adversarial attacks. *arXiv preprint arXiv:2310.10844*.

Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. 2020. Can you put it all together: Evaluating conversational agents' ability to blend skills. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2021–2030, Online. Association for Computational Linguistics.

Pei-Hao Su, Chuan-Hsun Wu, and Lin-Shan Lee. 2015. A recursive dialogue game for personalized computer-aided pronunciation training. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(1):127–141.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Gladys Tyen, Mark Brenchley, Andrew Caines, and Paula Buttery. 2022. Towards an open-domain chatbot for language practice. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 234–249, Seattle, Washington. Association for Computational Linguistics.

P. Vlugter, A. Knott, J. McDonald, and C. Hall. 2009. Dialogue-based CALL: a case study on teaching pronouns. *Computer Assisted Language Learning*, 22(2):115–131.

Jiashuo Wang, Haozhao Wang, Shichao Sun, and Wenjie Li. 2023. Aligning language models with human preferences via a bayesian approach. *arXiv preprint arXiv:2310.05782*.

Zhihui Zhang and Xiaomeng Huang. 2024. The impact of chatbots based on large language models on second language vocabulary acquisition. *Heliyon*, 10(3).

A Study setup details

Screening Our participants are recruited from Prolific and filtered using the built-in screening process. Participants must have a non-English language for their first language, primary language, and earliest language in life. As this does not guarantee that each participants' first language is *not* English (one can have multiple first languages), we also ask for their first languages later in the study. Additionally, we filter out participants living in countries where English speakers are in the majority (e.g. US, UK, Australia, etc.).

All participants' first languages can be found in Table 2.

First language	Number of participants
Polish	60
Portuguese	32
Italian	17
Greek	11
Spanish	11
Hungarian	8
German	7
Russian	3
Czech	3
Slovene	2
Afrikaans	2
Latvian	1
French	1
Arabic	1
Romanian, Moldovan	1
Urdu	1
Dutch	1
Turkish	1
Tagalog	1
Ukranian	1

Table 2: All first languages among our participants. Note that each participant can specify more than one first language.

Payment Before the study begins, participants are told that they will be paid a minimum of £7 for roughly half an hour of their time, including at least 15 minutes of chatbot interaction. Pay will increase with every additional 15 minutes spent with the chatbot(s), up to a maximum of £13. All entries are manually verified before payment to remove low-effort or invalid entries.

Consent form Participants are redirected to our website for the study, where they are presented with a consent form detailing how their data will be used. The consent form was written with second language speakers in mind, to ensure that beginner learners can also understand it. Participants can also contact the authors via email or the messaging system on Prolific regarding any concerns about the study. To proceed to the next section, participants must consent to their data being used for research purposes. However, they can withdraw their consent at any point, up to 6 months after the study. They may also exit the task any time they wished.

Profiling questions There are two questions in this section:

1. *What is/are your first language(s)?*

Participants can select one or more languages out of a list of ISO-639 languages.

2. *How long have you been learning English?* Participants enter a number followed by a choice of "years" or "months".

Proficiency questions 25 multiple choice questions were used to estimate the proficiency level of users. Questions are taken from the Cambridge English Test Your English application (General English) [4]. Participants are asked to select one of 3 or 4 options for each question. Scores are then converted to CEFR levels, as done on the website. This CEFR level is used as input to the difficulty adaptation mechanism (Tyen et al., 2022).

Chatbot interaction At the beginning of this section, participants are informed that:

1. They should not reveal any personal information, even if asked.
2. The chatbots are not real people, despite what the messages may say, but messages will be read by researchers afterwards.
3. There is a risk that the chabots may generate inappropriate messages. Participants can flag messages as inappropriate by clicking on the 'Flag as inappropriate' button. Clicking on the button again un-flags the message.
4. Information or opinions in the generated messages should not be taken for fact.
5. If participants are finding the messages difficult, they can flag messages as too difficult by clicking on the 'Flag as too difficult' button. Clicking on the button again un-flags the message.
6. There will be attention questions in the next section, so participants should read messages carefully.
7. (For those in the dictionary lookup condition) Participants can click on words to look them up in the dictionary.

After acknowledging the above, participants may begin the chatbot interaction. In both reading and chatting conditions, messages are generated on-the-fly, using an NVIDIA Tesla V100 GPU.

[4] https://www.cambridgeenglish.org/test-your-english/general-english/

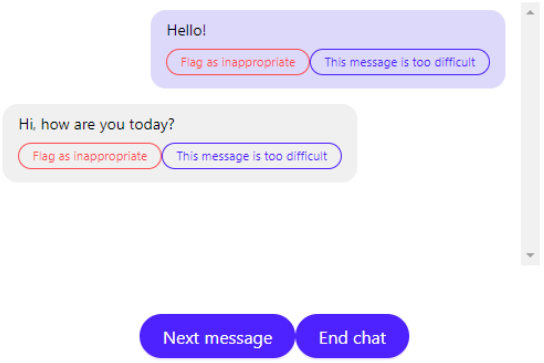

Figure 8: Interface presented to participants in the reading condition. Messages on both sides are chatbot-generated using the same parameters.

Reading condition In the reading condition, the user reads a conversation between two identical chatbots with the same settings. The user interface can be found in Figure 8. Unlike the UI for the chatting condition (in Figure 1), the user presses a button to reveal the next message, instead of typing in a text input field. Note that to maintain fair comparison, all messages in either the reading or chatting conditions are generated in real time.

Adaptive condition In the adaptive condition, all chatbot messages are generated using a weighted reranking decoding method (Tyen et al., 2022). This method consists of 3 components:

1. Sub-token penalties to adjust probabilities of tokens during generation
2. A reranker model to assign adjusted scores to each generated candidate message
3. A filter to remove generated candidates that contain ungrammatical words

For the reranker model, we use weights directly from https://github.com/WHGTyen/ControllableComplexityChatbot/tree/master/complexity_model without performing any additional fine-tuning. The final score of each generated candidate is calculated as the average rank between ranked probability scores and ranked complexity scores, weighting both equally:

$$\frac{r(P(C)) + r(|L_{\text{user}} - L_C|)}{2} \quad (1)$$

C is the candidate message; r is a ranking function returning a rank out of 20 candidates; L_{user} is the

CEFR level of the user, and L_C is the predicted CEFR level of the candidate message.

For the vocabulary filter, we use a list of English words from https://github.com/dwyl/english-words, but ignore capitalized words (indicating proper nouns). For the subtoken penalties, the probability of each token t is given by:

$$P(t) = \begin{cases} P(t) \cdot \varphi(L_t - L_{\text{user}}) & \text{if } L_t > L_{\text{user}} \\ P(t) & \text{otherwise} \end{cases} \quad (2)$$

where L_t refers to the CEFR level of token t and L_{user} refers to the user's CEFR level, determined by proficiency test scores at the beginning of the user study. The level is determined before any text is generated, does not change throughout the conversation, and is implemented in the same way regardless of reading/chatting or lookup conditions. For the function φ representing the normal distribution, we follow parameters used in the original paper, $\mu = 0$ and $\sigma = 2$.

Inappropriate language Participants have the ability to flag messages as being inappropriate. Of the 21,283 messages sent by a chatbot, 359 (1.69%) were flagged as such. A small sample reveals that about half of these messages were flagged due to being nonsensical, or logically or pragmatically unsuitable for the context, rather than offensive – this may be due to some participants misinterpreting the word "inappropriate". The remaining half generally touch on politically sensitive topics, use politically incorrect terms, or are offensive or insulting in some way.

The existence of these messages is concerning for chatbot usage in educational settings, especially for younger learners. Recent work on AI alignment has produced considerable improvements over the past few years (see Ji et al. (2023) for a comprehensive survey), but it is still possible to elicit inappropriate messages, especially when under specially crafted attacks (Shayegani et al., 2023). In its current form, we believe that LLM chatbots are best suited for an adult audience who are aware and informed of the nature of language models. However, current technology on LLM safety is improving rapidly, and new methods for mitigating toxicity are being developed constantly (e.g. Ouyang et al. (2022); Bai et al. (2022); Wang et al. (2023)), so it may soon be possible to deploy chatbots that are safe for younger audiences.

Feedback questions Table 3 shows the full list of questions asked after each chatbot session. Questions vary slightly depending on whether the participant is assigned the chatting or reading condition.

Questions (a) and (b) are attention questions used to eliminate low-effort entries where the participant failed to engage with the task. Among our submissions, only 4 are removed for this reason.

	Chatting condition	Reading condition
	Attention questions	
(a)	Were there messages from BlenderBot that did not make sense? If so, can you give some examples?	Were there messages from BlenderBot 1 or 2 that did not make sense? If so, can you give some examples?
(b)	Tell us one fact about BlenderBot that you learned from this conversation.	Tell us one fact about either BlenderBot 1 or 2 that you learned from this conversation.
	Likert questions	
(c)	I enjoyed chatting with BlenderBot.	I enjoyed reading the messages between BlenderBot 1 and BlenderBot 2.
(d)	I learned a new word/phrase while chatting with BlenderBot.	I learned a new word/phrase while reading these messages.
(e)	I learned new grammar while chatting with BlenderBot.	I learned new grammar while reading these messages.
(f)	I think I can learn more words / phrases / grammar if I chatted longer.	I think I can learn more words / phrases / grammar if I read more of these messages.
(g)	I feel more comfortable chatting with BlenderBot now than at the beginning.	N/A
(h)	Chatting with BlenderBot made me feel more confident about chatting in English with real people online.	Reading these messages made me feel more confident about chatting in English with real people online.
(i)	If I can chat with BlenderBot a few more times, it will make me feel more confident chatting online in English with real people.	If I can read more of these messages, it will make me feel more confident chatting online in English with real people.
(j)	I would like to continue using this chatbot in the future.	I would like to continue reading similar messages in the future.
	Feedback questions	
(k)	What did you think about the difficulty of the messages? Options: Too easy / A bit easy / Just right / A bit difficult / Too difficult	
(l)	Do you have any other thoughts, comments, or feedback for us? (free text response)	

Table 3: Questions answered by each participant after their chatbot session.