

# Curriculum-Driven Edubot: A Framework for Developing Language Learning Chatbots Through Synthesizing Conversational Data

YU LI\*, Columbia University, USA

SHANG QU\*, University of Science and Technology of China, China

JILI SHEN, Zhejiang University, China

SHANGCHAO MIN, Zhejiang University, China

ZHOU YU, Columbia University, USA

Chatbots have become popular in educational settings, revolutionizing how students interact with material and how teachers teach. We present Curriculum-Driven EduBot, a framework for developing a chatbot that combines the interactive features of chatbots with the systematic material of English textbooks to assist students in enhancing their conversational skills. We begin by extracting pertinent topics from textbooks and then using large language models to generate dialogues related to these topics. We then fine-tune an open-source LLM using our generated conversational data to create our curriculum-driven chatbot. User studies demonstrate that our chatbot outperforms ChatGPT in leading curriculum-based dialogues and adapting its dialogue to match the user’s English proficiency level. By combining traditional textbook methodologies with conversational AI, our approach offers learners an interactive tool that aligns with their curriculum and provides user-tailored conversation practice. This facilitates meaningful student-bot dialogues and enriches the overall learning experience within the curriculum’s pedagogical framework.

CCS Concepts: • **Human-centered computing** → **Human computer interaction (HCI)**; **Collaborative and social computing**; **Interaction design**.

Additional Key Words and Phrases: Dialogue Systems, Education, Large Language Models, Language Learning

## ACM Reference Format:

Yu Li, Shang Qu, Jili Shen, Shangchao Min, and Zhou Yu. 2023. Curriculum-Driven Edubot: A Framework for Developing Language Learning Chatbots Through Synthesizing Conversational Data. In *CHI’24, May 11–16, 2024, Honolulu, Hawaii*. ACM, New York, NY, USA, 28 pages. <https://doi.org/XXXXXXX.XXXXXXX>

## 1 INTRODUCTION

The emergence of conversational agents has had a significant impact on educational technology, changing the way students interact with material and how teachers impart knowledge [12, 25, 41]. These agents, more commonly known as “chatbots,” have shown their usefulness in a range of educational settings, from teaching computer programming [11] to strengthening conversational skills [23]. However, its broad application comes with inherent challenges, especially in the area of conversational skill development. Many chatbots primarily answer users’ questions and follow given instructions. However, language learning is typically organized and follows a textbook curriculum. As learners work through these materials, they expect logical order and consistent content. Unfortunately, common chatbots, with their

\*Both authors contributed equally to the work.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

© 2023 Association for Computing Machinery.

Manuscript submitted to ACM

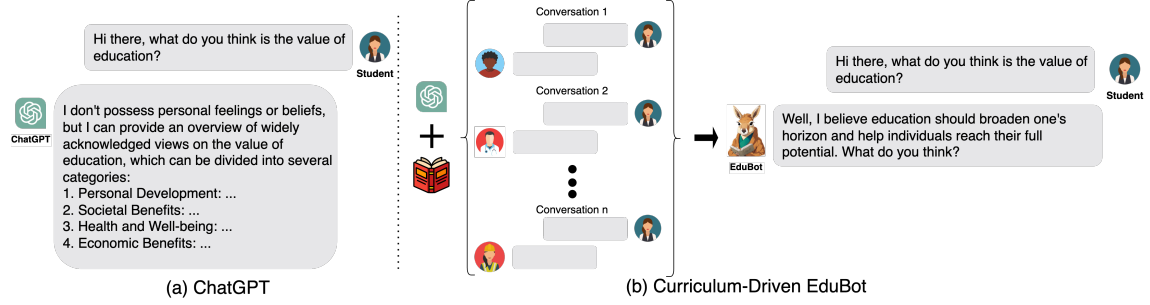


Fig. 1. Comparison between ChatGPT vs. Our Curriculum-Driven EduBot. ChatGPT operates as an AI-powered Q&A tool, delivering comprehensive responses from a broad knowledge base. The Curriculum-Driven EduBot is fine-tuned with synthesized conversations, offering an interactive and adaptive learning experience through conversational practice.

generic conversations, may present language or content that are unsuitable for a student’s level of knowledge, which can hinder their learning progress.

To address these challenges, we propose a framework, Curriculum-Driven EduBot, to develop a chatbot based on a curriculum. This chatbot will focus on the given topics and use vocabulary words from the curriculum to match the users’ English proficiency. It will act as a conversational practice partner, combining the interactive features of chatbots with the organized content of English textbooks. Specifically, we start by extracting relevant topics from textbooks and utilizing large language models (LLMs) to synthesize fixed-format personas to represent both participants in the dialogue. Subsequently, we use LLMs to synthesize dialogues around these topics and personas while including the vocabulary found in the textbook. Afterward, we fine-tune an open-source LLM with our generated conversational data to construct our chatbot. Our chatbot is not just a responsive tool, but an academic companion that guides students through coherent and friendly dialogues tailored to their English proficiency level. As illustrated in Figure 1, existing chatbots, such as ChatGPT, are not based on a curriculum. Instead of being conversational learning partners, they mainly act as AI-driven Q&A systems, and their content may not always be in line with the student’s educational objectives. On the contrary, our bot is constructed on synthesized dialogues that include clearly defined characters, curriculum-appropriate topics, and textbook-based vocabularies, thus providing an interactive and user-tailored conversational experience.

We empirically assessed our Curriculum-Driven EduBot using a high-quality college English textbook designed for English learners. We conducted an intensive user study with a group of students. The comparative evaluations show that our bot outperforms ChatGPT in various metrics. 75% of students believed that EduBot was especially adept at leading conversations to encourage interactive practice, and students generally thought it was better suited to their English proficiency than ChatGPT. The results and conversation examples from the user study clearly demonstrate that our bot is more closely aligned with the role of a language-learning companion. 83.3% of students are willing to recommend EduBot to others. 87.5% of students believe that interactions with EduBot can help students improve their conversational skills. In summary, our main contributions are as follows:

- We introduce a curriculum-driven chatbot framework. First, we synthesize dialogues that incorporate fixed-format personas, curriculum topics, and related vocabularies. We then fine-tune an open-source LLM to implement the chatbot, seamlessly merging chatbot interactivity with structured educational material.
- We applied our framework to a specific curriculum. User studies reveal that our EduBot outperforms existing state-of-the-art systems. 87.5% of students believe that EduBot can help them improve their conversational skills.

## 2 RELATED WORK

Many studies have shown that Artificial Intelligence (AI) can be utilized in educational settings. [6, 10, 14]. For example, Rodrigues and Oliveira [31] created a formative assessment system that is capable of creating and assessing tests, as well as tracking the progress of the learners. Similarly, Lan et al. [22] proposed a machine learning-based approach to learning analytics, highlighting its potential to assess student knowledge. Recent advances in LLM [2, 19, 26, 27, 33] have had a major impact on the use of chatbots in educational settings [12, 21, 25]. These conversational agents provide personalized learning experiences, engage learners, and help them retain knowledge. For example, Vasconcelos and dos Santos [38] investigated the capabilities of ChatGPT<sup>1</sup> and Bing Chat<sup>2</sup> as resources that foster critical thinking and understanding of concepts to improve STEM education. Moreover, Li et al. [23] used chatbots as conversational practice partners, providing learners with automatic grammar error feedback for language learning. Building on these advances, our research takes advantage of the most advanced open-source chatbot, offering students the opportunity to engage in conversations that are in line with their curriculum.

Language learning, traditionally dependent on static resources such as textbooks and structured courses, has been greatly enhanced by curriculum-aligned approaches that combine consistency with adaptability. Krashen [20] highlighted the importance of customized content delivery in language learning, suggesting that when learners engage with material that aligns with a structured curriculum, they often experience better comprehension and retention. Many researchers advocate for systematically integrating curriculum content into newer learning platforms to provide contextually relevant language exposure [1, 13, 24]. For example, Rodríguez-Castro [32] explored the potential of digital tools, such as virtual reality simulation, that map their content to official language learning curricula, ensuring that learners stay on track while taking advantage of interactive digital experiences. Furthermore, Ho et al. [15], Holden and Sykes [16] have demonstrated the power of curriculum-based gamification in language learning. By connecting game elements with curriculum milestones, they have shown that it can motivate learners and encourage them to stay engaged for longer. Qian et al. [29] applied lexically constrained decoding to a dialog system, encouraging it to include words and phrases that were aligned with the curriculum. This resulted in better comprehension of target words by students and increased interest in practicing English. Subsequent to these efforts, our chatbot is the first of its kind to synthesize conversations from curricula and train on an open-source LLM.

The use of pre-trained language models (PLMs) [4, 26, 28, 30, 34, 35, 39, 42] has enabled the generation of synthetic conversational data to enrich limited datasets, particularly in privacy-sensitive domains such as the medical domain [37]. Research has been conducted to use PLMs to augment a variety of conversational datasets [7–9, 17, 43]. For example, Zheng et al. [43] and Chen et al. [8] used GPT-J [39] to generate responses tailored for emotional support dialogues and comprehension tasks, respectively. Kim et al. [18] proposed a collaborative human-AI paradigm in which a human operator and GPT-3 alternate in conversation. Chen et al. [7] generated both dyadic and multiparty dialogues grounded on specific topic words, demonstrating that their outputs were of comparable quality to those crafted by humans. Our approach focuses on the generation of comprehensive conversations derived from educational curricula, allowing us to control personas, topical focus, and lexical choices throughout the synthetic data creation process.

## 3 METHOD

We propose a framework for building a curriculum-based chatbot that could converse on topics derived from a given curriculum while aligning its responses to the user’s English proficiency level. As shown in Figure 2, our development

<sup>1</sup><https://chat.openai.com>

<sup>2</sup><https://www.bing.com/new>

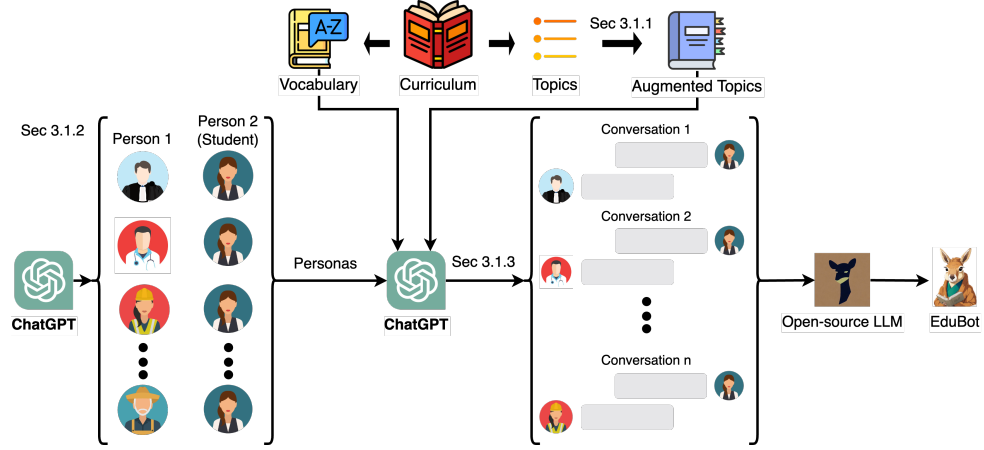


Fig. 2. The initial step of the Curriculum-Driven EduBot Development is to enhance textbook topics (Sec.3.1.1). Following this, personas are created for synthetic conversation participants (Sec.3.1.2). Then dialogues are constructed based on vocabulary, topics and personas (Sec. 3.1.3). After this, an open-source model Vicuna is fine-tuned to get the EduBot ready for deployment (Sec. 3.2)

process is divided into two parts. First, we use ChatGPT to generate simulated human-to-human dialogues based on textbook topics. Then, we fine-tune an open-source LLM to create our chatbot.

### 3.1 Conversational Data Augmentation

The art of synthesizing human-human dialogues is largely dependent on two key elements: the topics being discussed and the personas of the dialogue participants [7, 8, 17]. To synthesize dialogues based on a curriculum, we propose a three-step procedure. We start by extracting the main topics from the textbook and generating associated subtopics. Second, we design a range of personas for the participants of the synthetic dialogues. Last, we synthesize dialogues based on the topics and personas obtained in the previous steps.

**3.1.1 Augment Topics.** The range of topics covered in each unit of a curriculum is often limited. To broaden our synthetic dialogues to include a wide range of topics, we first extract the primary topics of the curriculum and then use ChatGPT to generate associated subtopics for each primary topic. For example, in our application, the primary topic of the first unit is “The True Value of Education”. We expand it to more topics such as “The importance of education in personal and professional development” and “The role of education in promoting social justice and equity”. This process ensures that our dialogues are comprehensive and varied. We steer this augmentation process using specific prompts provided to ChatGPT:

- Given an input topic, generate a list of <n> closely related topics that could be explored further.

Input topic: <Topic>

Further information on this step, along with sample input-output pairs, can be found in Appendix A.1.

**3.1.2 Create Personas.** To enrich the conversational context, we also prompt ChatGPT to create personas for two dialogue participants: Person 1 and Person 2. These personas are crafted to reflect diverse backgrounds, including demographic characteristics (e.g. gender and race), socioeconomic status, cultural distinctions, Myers-Briggs Type

Indicator (MBTI) personality profiles, and personal experiences. Since the dialogue occurs between our chatbot and a student, and the model is trained to take on the role of Person 1 in the dialogue, we specify that Person 2’s background information consistently represents a student, ensuring that their conversational role aligns with that of a typical student for the textbook we choose. In contrast, we randomly generate Person 1’s background information. Adopting this ‘fixed-random’ strategy offers two primary benefits: 1. It enables our chatbot to be trained with the student persona acting as the user and the alternate persona as the chatbot. Thus, the chatbot is ready to anticipate that its user is a student. 2. This encourages ChatGPT to generate conversations about topics that are commonly discussed by students, such as college life, which increases the bot’s appeal to users from this background. A detailed description of the prompt structure for this step is as follows:

- Please provide me with one individual Person 1 with different backgrounds, including information about their demographic, socio-economic status, culture, MBTI personality type, and personal experiences, no need to show names. Then provide me with one individual Person 2 who is a <student role information> but with different information.

We can substitute the <student role information> with a comprehensive and detailed description of the students who actually utilize the textbook we select. More information about this step, along with an example of input and output, can be found in Appendix A.2.

**3.1.3 Compose Dialogues.** Using the generated personas and topics, we now instruct ChatGPT to generate synthetic dialogues. To tailor the dialogue to the user’s English proficiency level and ensure that the dialogue aligns with the vocabulary that students are familiar with, we follow Qian et al. [29] and extract a subset of words from the vocabulary list of the relevant textbook unit to integrate into the conversation. We instruct ChatGPT to use a pair of personas generated in Step 2, one fixed as a student, and the other with randomized characteristics. Participants with these personas will use the words in the vocabulary and converse on a topic chosen from our extended topic list in Step 1. To help users engage in the conversation and have a better user experience, we also follow previous work and instruct the chatbot to actively lead the dialogue [36]. Person 1, which represents the chatbot in these synthetic dialogues, is prompted to guide the dialogue. This deliberate configuration allows our chatbot to take the conversational lead with students, providing direction and guidance. The prompt given to ChatGPT in this step is provided below.

- Generate a single conversation between these two people as Person 1 and Person 2 about the topic <Topic>.  
Please take into account their distinct personalities and their backgrounds. Begin the conversation with Person 1.  
Please include the following keywords in Person 1’s utterances: <Vocab>  
Person 1 should guide the conversation by asking more questions.

More details about this step, as well as examples of input and output, are provided in Appendix A.3.

### 3.2 Fine-Tuning An Open-Source Language Model With Synthesized Conversational Data

After obtaining the synthesized dialogues, we use them to fine-tune an open-source large language model. Utilizing such open-source models offers several advantages: First, we can take advantage of the underlying architecture and parameters of open-source models to customize them to our needs. In addition, we can synthesize additional data

as needed and improve the model through successive iterations. Lastly, open-source models are usually free, which significantly reduces costs.

For our particular application, we choose Vicuna-13B<sup>3</sup>, a state-of-the-art open-source language model. We use it to build our chatbot since it possesses impressive understanding capabilities and demonstrates comparable performance to ChatGPT. We fine-tune a single Vicuna-13B model using topics taken from all the units in the textbook. This approach ensures that our chatbot has a comprehensive knowledge base for all topics in the textbook. During training, the chatbot takes on the role of Person 1, while the student takes on the role of Person 2. Following [3], we design the prompt structure for Vicuna as follows:

- As a social chatbot, please engage in a conversation while adopting the following personas:  
 <Person 1 Persona>.  
 Engage in a conversation about <Topic> by showcasing your personas.  
 Share interesting anecdotes, facts, and experiences related to <Topic>  
 The English level of the conversation should be at CEFR <English Proficiency Level of Textbook>.

To ensure that our bot is compatible with the English proficiency level of the textbook, we use The Common European Framework of Reference for Languages (CEFR) to control the difficulty level of language in our training process. CEFR is a widely used method to classify the difficulty level of texts. It defines six levels that represent increasing levels of difficulty or proficiency: A1, A2, B1, B2, C1 and C2. We include the CEFR level of the textbook in our system prompt. More information on our implementation can be found in Appendix B.

### 3.3 Deployment of the Fine-tuned EduBot

When a student uses EduBot, they first chooses a particular unit from the curriculum. We then randomly assign a persona to the chatbot, pick a topic from the augmented topic list of the chosen unit, and randomly sample a set of words from the “new words” vocabulary of the unit. We incorporate the above information in a specialized prompt which is structured as follows:

- As a social chatbot, please engage in a conversation while adopting the following personas:  
 <Persona>.  
 Engage in a conversation about <topic> by showcasing your personas.  
 Share interesting anecdotes, facts, and experiences related to <Topic>.  
 Include the following words in your utterances: <Vocab>.  
 The English level of the conversation should be at CEFR <Textbook English Proficiency>.

EduBot’s deployment prompt is similar to its training prompt, except it incorporates the vocabulary of new words. This consistency helps EduBot generate responses related to the topic, draw on the persona’s background information, and keep the language difficulty level consistent, just like Person 1 in the synthetic dialogues. Additionally, by including new words, EduBot helps students to revisit terms in the corresponding material. Examples of dialogues between EduBot and students can be found in Appendix F.

<sup>3</sup><https://lmsys.org/blog/2023-03-30-vicuna/>

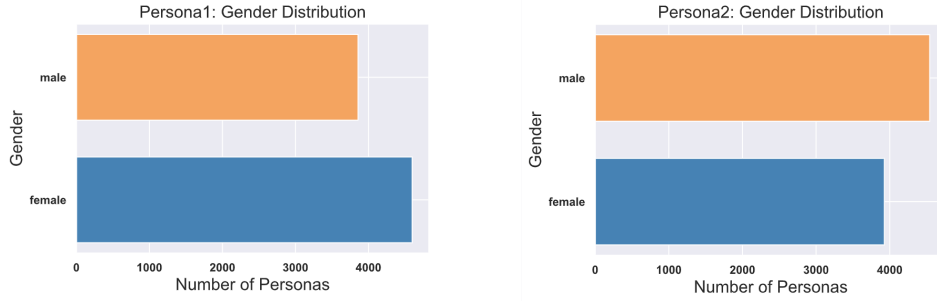


Fig. 3. Distribution of gender in personas

#### 4 CURRICULUM SOURCE

To evaluate our method and customize our chatbot for students interested in improving their conversational skills, we choose “New College English” (3rd edition), a textbook widely used in Chinese universities. This series has four levels of proficiency, each with a “Comprehensive Tutorial” and an “Audiovisual Said Tutorial”. Since we want to enhance students’ conversational skills, we choose the “Audiovisual Said Tutorial” from the third level, which is the material used in the most advanced English course. This tutorial consists of 8 units, each with a list of conversation topics. An example of a topic from Unit 1 is “The True Value of Education”. These predetermined topics are the basis for our topic augmentation. For each main topic, we create 10 associated topics through the process outlined in Section 3.1.1. For instance, a possible associated topic for “The True Value of Education” is “The importance of education in personal and professional development”. Additionally, since each unit introduces a set of new words, we randomly select 10 words from the word list of the corresponding unit to incorporate into each dialogue composition, as described in Section 3.1.3. Through this approach, we are able to generate a total of 7,687 dialogues across the 8 units, in preparation for the subsequent development stages.

#### 5 CONVERSATIONAL DATA STATISTICS

Using our chosen curriculum as the basis, we synthesized 880 to 1,210 dialogues per unit, averaging 1,058.76 dialogues for each. These dialogues comprise an average of 11.77 utterances, on average containing 28.71 words each. This section analyzes the statistical characteristics of our synthesized dialogues. To ensure the quality of our conversation data and its alignment with our objectives, we employed three attributes in our data synthesis process: curriculum topics, fixed-format personas, and relevant vocabularies. We first examine our generated personas for diversity and breadth in Sec. 5.1. Then we evaluate the distribution of target words within dialogues in Sec. 5.2. Moreover, to ascertain the congruence of our dialogues with the English proficiency standards of the textbook, we leveraged ChatGPT to assess word difficulty levels in both our synthesized dialogues and the curriculum in Sec. 5.3.

##### 5.1 Persona Trait Distribution

As elaborated in Section 3.1.2, including conversation personas is important for ensuring diverse, engaging conversation content and styles. We first examine the range of personality traits represented in the generated personas. We use keyword string matching to extract the persona trait values from the generated persona descriptions. Figures 3 and 4 show the gender and MBTI personality type distributions of the personas, respectively. Synthetic dialogues include nearly

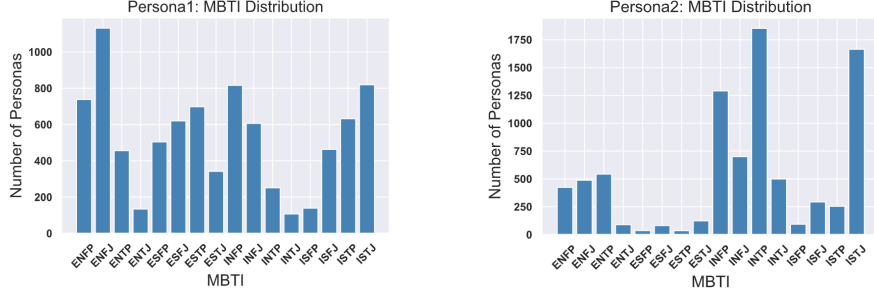


Fig. 4. Distribution of MBTI personality types in personas

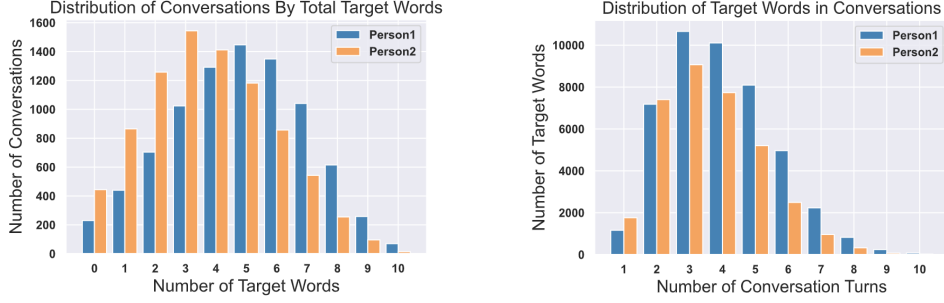


Fig. 5. Distribution of target words

equal proportions of both genders. The personality type distribution is not uniform, but all 16 types are represented in the synthetic dataset.

In addition, we verify the nationalities in the persona descriptions of Person 2. 8,000 of the total 8,470 persona descriptions explicitly specify “China” or “Chinese”. This indicates that in most cases, ChatGPT successfully followed the additional instructions regarding Person 2, mentioned in Section 3.1.2.

## 5.2 Target Word Distribution

During synthetic conversation generation, we included 10 target words in each prompt to be included in Person 1’s utterances. Therefore, for each synthetic dialogue created, we compute the number of times the target words in the prompt are used in each dialogue turn. The first graph in Figure 5 displays the distribution of dialogues based on the total number of target words included by Person 1 and Person 2, respectively. Most of the words are included in Person 1’s utterances, and in the majority of dialogues, Person 1 mentions at least half of the 10 vocabulary words. The second graph in Figure 5 shows the total number of vocabulary words included in each dialogue turn for each person.

## 5.3 English Proficiency Level

We evaluate whether the English proficiency level of the generated dialogues is similar to that of the curriculum. We use ChatGPT as an evaluator, as it has demonstrated its prowess in various language evaluation tasks Chang et al. [5], Wang et al. [40], Zheng et al. [44]. We follow Zheng et al. [44] and utilize ChatGPT to automatically classify dialogues according to the CEFR scale using the following prompt:



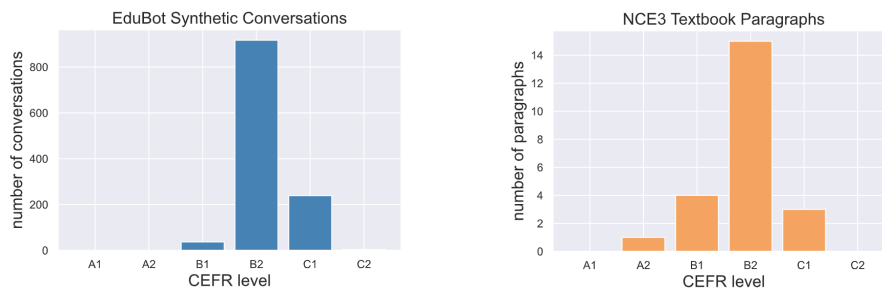


Fig. 6. English proficiency levels of synthetic conversations and textbook paragraphs

- Evaluate the English proficiency of the given conversation according to the CEFR scale.

Provide one of the following six answers: A1, A2, B1, B2, C1, C2.

Output the CEFR level of the following conversation: <conversation>

<conversation> corresponds to the complete synthetic dialogue to be evaluated.

We then use the same method to evaluate the English proficiency level of “New College English” (3rd Edition), the original textbook we choose, by replacing the last sentence of the prompt with:

- Output the CEFR level of the following paragraph: <paragraph>

We assess each paragraph in the sample texts from “New College English”. The results of our evaluation for Unit 1 are shown in Figure 6. We found that synthetic dialogues are comparable to those found in textbooks, yet they are slightly more challenging. This indicates that our method of synthesizing dialogues effectively ensures that our dialogues match the English proficiency level of the original textbook.

## 6 EXPERIMENTS

### 6.1 Baseline

To assess our chatbot’s performance and the effectiveness of our training approach, we use ChatGPT as our baseline. ChatGPT is a state-of-the-art conversational agent that is well known for its ability to generate meaningful and contextually appropriate conversations. It has been trained on a variety of datasets, allowing it to respond to a wide range of topics. We prompt ChatGPT to act as a chatbot that can chat about primary textbook topics with students. We do not employ zero-shot prompted Vicuna as our baseline because it frequently fails to adhere to our instructions. The dialogue utterances it produces are often lengthy and difficult to comprehend, making it impractical to conduct a conversation with the chatbot. This is due to the smaller size and weaker instruction-following capability of Vicuna in comparison to ChatGPT. Our fine-tuning approach improves the Vicuna model and resolves the aforementioned issue.

In addition, we note that the length of the responses has a major effect on user experience. Some students preferred longer responses given by the chatbot. This could be due to the text-based format of our chatbot. Compared with using speech-based chatbots, users may be more accepting of long responses when using text-based chatbots, since the redundancy of responses is less pronounced. However, lengthy replies from the chatbot might hinder the development of the user’s conversational skills since they may be more inclined to simply read the material and provide short responses instead of proactively engaging in the conversation. To ensure that the assessment of conversational skills is fair, we restrict the length of ChatGPT responses. The prompt we use for ChatGPT is as follows:

- As a social chatbot, please engage in a conversation about <Topic>. Share interesting anecdotes, facts, and experiences related to <Topic> Each response should be either one or two sentences. Please make all responses short and concise. Follow the above rules for all your utterances.

While our prompt effectively encourages succinct responses from ChatGPT, we observed occasional cases in our user study where ChatGPT ignored the brevity instructions and produced lengthy replies. This exception might be attributed to certain types of user input, such as when a user requests an in-depth explanation.

## 6.2 Experimental Settings

**6.2.1 Participants.** We recruited 24 participants in the user study from a renowned university in China through student discussion forums and in class. All participants were students from the university who have taken “College English 4”, the course corresponding to the “New College English” (3rd edition) textbook, within the past year. To register for the experiment, participants were required to submit a background survey.

In total, 48 students completed the background survey, 24 of whom (4 males and 20 females) completed the entire experiment and provided valid results. 19 of the participants were in their second year of university, 4 in their third year, and 1 in their fourth year. The average age of the participants was 19.26 years. The participants in the user study came from 20 different majors and had a variety of English proficiency levels. The total amount of time they had spent learning English ranged from 8 to 15 years, with an average of 11.65 years, and their final grades for “College English 4” ranged from 2.1/5.0 to 5.0/5.0, with an average of 4.06/5.0.

**6.2.2 Procedures.** We performed experiments in which participants were assigned either Unit 1 or Unit 2 of the textbook. Each participant had two conversations with both EduBot and ChatGPT, with each conversation containing a minimum of 20 utterances. To avoid bias from the naming and order of experiments, we randomly labeled the two bots A and B for each session and asked the participant to converse first with Bot A and then with Bot B.

Participants completed a questionnaire immediately after interacting with the two chatbots. First, they were asked to summarize each of their four conversations with the chatbots. The main part of the questionnaire consists of 20 criteria divided into 6 categories: Consistency with the curriculum, English proficiency level, role identification, quality of the conversation language, quality of conversation content, and general usefulness. For each criterion, participants had to choose between 3 options: Bot A was better, Bot B was better, or the two bots were the same. All questions and instructions were provided in both Chinese and English, and participants were encouraged to refer to their conversation records and textbook content while completing the questionnaire. Each study took between 20 and 30 minutes, and participants were given \$5 as compensation, which is in line with the minimum wage standards in China<sup>4</sup>. We excluded one submission where the student did not summarize the dialogues correctly and three submissions that contained self-conflicting answers.

In Appendix D, we present the user interface used in our experiments. The complete background survey and questionnaire are provided in Appendix C.1 and Appendix C.2.

Table 1. Questionnaire Results (Percentage of Responses)

Section	Question	EduBot (%)	ChatGPT (%)	Same (%)
Consistency With Curriculum	1. The main topics of my conversations with the chatbot were closely related to what I learned in English class.	41.7	50.0	8.3
	2. The chatbot brought up anecdotes, examples, questions, etc., related to what I learned in English class.	25.0	41.7	33.3
	3. The chatbot mentioned topics and content that were not directly covered in the textbook and course.	50.0	16.7	33.3
English Proficiency Level	1. During our conversations, the chatbot mentioned some vocabulary words that I learned in my English course.	37.5	37.5	25.0
	2. The chatbot used many vocabulary words that I didn't understand.	20.8	37.5	41.7
	3. I didn't find the conversations too easy to be helpful.	16.7	29.2	54.2
Role Identification	1. During conversations, I felt that the chatbot recognizes that I am a Chinese college student.	41.7	29.2	29.2
	2. During the two conversations with the chatbot, I felt like I was talking with two different people.	20.8	12.5	66.7
Language Quality	1. The utterances provided by the chatbot were coherent and fluent.	29.2	25.0	45.8
	2. The chatbot's responses were concise and accurate.	50.0	12.5	37.5
	3. Unlike in real everyday conversations, the chatbot's responses were long and redundant at times.	8.3	66.7	25.0
	4. Interactions with the bot were similar to natural, realistic conversations and not overly formal.	62.5	4.2	33.3
Content Quality	1. The chatbot acknowledged what I said and provided reasonable responses.	37.5	41.7	20.8
	2. The chatbot provided unique and personal perspectives regarding the selected topic.	45.8	37.5	16.7
	3. The chatbot used personal experiences to support its opinions.	33.3	16.7	50.0
	4. The chatbot actively raised questions to guide the course of the conversation.	75.0	4.2	20.8
	5. The chatbot didn't output offensive or hurtful responses.	0.0	8.3	91.7
General Usefulness	1. I would find it useful to use the chatbot to review what I learned in class.	16.7	25.0	58.3
	2. I would recommend the chatbot to other students.	37.5	16.7	45.8
	3. I believe that continuing to use the chatbot will help me improve my English conversation skills.	25.0	12.5	62.5

## 7 RESULTS AND DISCUSSION

### 7.1 Results

The full results of the user study are shown in Table 1. We show the win rates corresponding to each criterion in the questionnaire. Results show that EduBot outperforms ChatGPT in several different aspects.

*7.1.1 EduBot’s language quality was on par with ChatGPT.* Similar percentages of participants preferred EduBot (29.5%) and ChatGPT (25.0%) regarding the coherence and fluency of the chatbots’ utterances. This shows that EduBot is able to produce responses with high language quality. It is remarkable that Vicuna-13B did not perform well without our fine-tuning, thus demonstrating the considerable improvements that our synthetic data enabled.

*7.1.2 EduBot offers a diverse range of relevant dialogue topics.* Through topic augmentation based on the curriculum, we aim for EduBot to center its conversations around topics that are relevant but not directly covered in the textbook. Significantly more participants chose EduBot (50.0%) over ChatGPT (16.7%) when asked which chatbot mentioned topics and content that were not directly covered in the textbook and course. This shows that EduBot is capable of discussing a wide range of diverse topics, compared with ChatGPT, which was only prompted with topics taken directly from the textbook.

At the same time, EduBot’s conversation content remains in line with the curriculum. When asked which chatbot’s conversation topics are more related to the “College English 4” course, student opinions were almost evenly divided. EduBot does not perform as well as ChatGPT in bringing up anecdotes, examples, questions, etc., related to the course. We believe that this is because ChatGPT gives longer statements that provide more material, while EduBot’s answers are more concise and concentrated on inquiring and engaging the user. This contrast is discussed in greater detail in Section 7.1.4.

*7.1.3 EduBot’s conversations align better with students’ English proficiency levels.* The same percentage of participants (37.5%) chose EduBot and ChatGPT regarding which chatbot provided more vocabulary that they learned in their English course. We believe this is because outputs produced by ChatGPT are generally close to CEFR B2 level without additional guidance, which makes it difficult to highlight EduBot’s alignment with the students’ English proficiency level. For the second question, 37.5% students claimed that ChatGPT used many vocabulary words that they did not understand, compared with 20.8% who chose EduBot. This shows that ChatGPT’s conversations were sometimes too difficult for the target users in our experiment.

An important reason for the relatively mixed results in this section was the variance in English proficiency levels among user study participants. We investigate the different preferences of students with different English levels in Appendix E.

*7.1.4 EduBot’s conversations are more natural and realistic.* The participants of the user study found that their conversations with EduBot were more natural and similar to real-life interactions. Such a distinction was caused by EduBot having access to synthetic dialogues that emulated real-life conversations of Chinese college students during the fine-tuning stage. A higher percentage of students thought that EduBot was concise and accurate (50% compared to 12.5% for ChatGPT), natural and realistic (62.5% compared to 4.2% for ChatGPT). On the other hand, the majority of participants thought ChatGPT’s responses to be too long and repetitive. Furthermore, results show that EduBot was better at guiding the conversation. 75% of students agreed that EduBot asked questions to guide the conversation,

<sup>4</sup><https://take-profit.org/en/statistics/wages/china/>

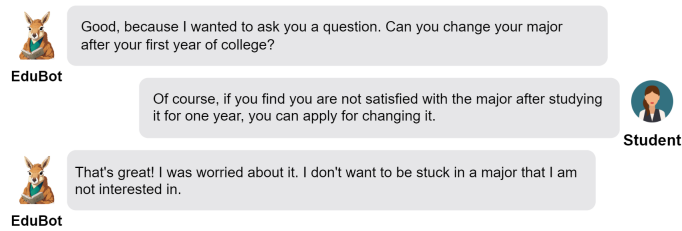


Fig. 7. Case 1

compared to only 4.2% who preferred ChatGPT. This means that when talking with EduBot, users can more easily follow the dialogue without having to come up with new topics to keep the conversation going.

**7.1.5 EduBot acknowledges the personas of both dialogue participants.** When conversing with EduBot, a larger proportion of participants felt that the chatbot was aware that they were Chinese college students (41.7%) compared to when they were talking to ChatGPT (29.2%). During the conversations, EduBot showed that it had knowledge of the user's identity as a Chinese college student by customizing its answers to the user's role. When the chatbot brought up content related to common experiences of college students, participants could easily build upon these topics and continue the conversation.

EduBot was also able to showcase its assigned personas during conversations in the user study. Slightly more participants felt like they were talking to two different people in their two conversations when using EduBot. EduBot performed better than ChatGPT in providing unique perspectives (45.8% compared to 37.5%) and personal experiences (33.3% compared to 16.7%) during conversations. The incorporation of diverse personas makes EduBot's conversations more realistic and engaging.

Examples of the above two characteristics are presented and further discussed in Section 7.2.

**7.1.6 EduBot is more useful for developing English conversation skills.** EduBot proved to be more advantageous than ChatGPT in two out of three criteria for general usefulness. A greater proportion of participants expressed their preference for EduBot in terms of recommending it to other students and using it to enhance their conversational skills. On the contrary, a slightly larger number of participants thought that ChatGPT would be beneficial for reviewing what they had learned in class. As discussed in previous sections, the probable explanation is that ChatGPT is more elaborate in presenting relevant information. Generally, EduBot is more suitable for our primary objective of developing the user's English conversation skills.

The results show that EduBot offers diverse conversation topics, takes on specific personas, acknowledges the persona of its target user, and provides natural and realistic utterances. On the other hand, the advantage of ChatGPT mainly lies in its ability to provide detailed and elaborate content surrounding the conversation topic. In general, EduBot is a more effective conversational partner that better cultivates the user's English conversation skills.

## 7.2 Case Study

In this section, we present several dialogue snippets from our user study to illustrate the benefits and drawbacks of EduBot.

**7.2.1 . EduBot recognizes the role of the user.** During our user study, EduBot identified that the user is a Chinese

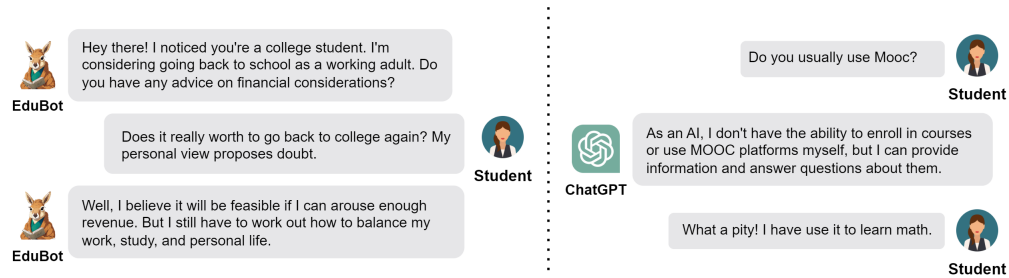


Fig. 8. Case 2 from EduBot (Left) and Case 3 from ChatGPT-Bot (Right)

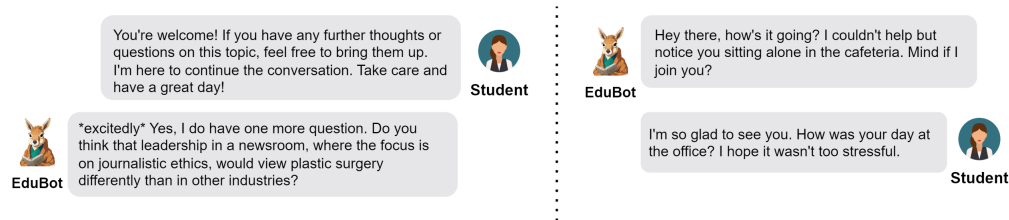


Fig. 9. Case 4 (Left) and Case 5 (Right)

college student. To make the conversation more engaging, it opened with topics or questions related to university life that the user was familiar with. For example, in case 1 (Figure 7), EduBot asked a question about the user's university policy, which naturally led the student into the conversation.

**7.2.2 EduBot provides personal opinions and experiences.** EduBot was capable of assuming the persona assigned in the prompt and then formulating its own opinions and experiences that are consistent with the persona. This made the conversations with users more realistic and engaging. In case 2 (Figure 8), EduBot took on the persona of a working adult and provided personal experience related to the given topic (continuing education after starting work). In contrast, ChatGPT usually did not provide a realistic answer when users asked about their personal experiences, which disrupted the natural flow of the conversation. An example of this is the dialogue in case 3 (Figure 8), which is about the benefits of MOOC, an online course platform. Unfortunately, ChatGPT was unable to provide a suitable response, as it has not been specifically trained to adopt personas and provide personal experiences.

**7.2.3 Limitations of EduBot.** We observe the following two phenomena that limited the quality of EduBot's conversations in several user study cases. First, EduBot occasionally included descriptions of its emotions or actions which should not appear in normal conversations. An example is given in case 4 (Figure 9). Such comments were included in some of the synthetic dialogues, causing the fine-tuned EduBot to generate them in its conversations.

Second, EduBot sometimes made incorrect assumptions about the user's feelings or the context of the conversation, then referred to them at the start of the conversation. For example, in case 5 (Figure 9), EduBot hallucinated that the user was alone in the cafeteria. This was due to ChatGPT hallucinating these scenarios in the generated data, which was then incorporated into our chatbot.

In the future, to address these two issues, we plan to further refine our data synthesis process and include stricter post-processing methods to filter out unnatural content.

## 8 CONCLUSION AND FUTURE WORK

In this work, we present a framework, Curriculum-Driven EduBot, to develop a curriculum-based chatbot, which combines the structured nature of English textbooks with the dynamic nature of chatbot interactions. We extract relevant topics from textbooks and then use large language models to synthesize conversations around these topics. We then fine-tune an open source LLM using these conversational data. Our user studies show that EduBot is more effective than platforms like ChatGPT in facilitating curriculum-related discussions, and is also able to adjust chatbot to match the user's English proficiency. These results demonstrate EduBot's ability to go beyond mere interactivity, providing learners with a contextually appropriate conversational platform that facilitates the development of conversation skills. In the future, there are opportunities to expand the content spectrum, incorporate multimedia elements, and introduce real-time feedback mechanisms. As we harness these enhancements, our vision is to see EduBot evolve into an indispensable learning companion.

## REFERENCES

- [1] Maria de Lourdes Andrade. 2014. Role of technology in supporting English language learners in Today's Classrooms. (2014).
- [2] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. Constitutional AI: Harmlessness from AI Feedback. *arXiv:2212.08073* [cs.CL]
- [3] Jianzhu Bao, Rui Wang, Yasheng Wang, Aixin Sun, Yitong Li, Fei Mi, and Ruifeng Xu. 2023. A Synthetic Data Generation Framework for Grounded Dialogues. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Toronto, Canada, 10866–10882. <https://doi.org/10.18653/v1/2023.acl-long.608>
- [4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1877–1901. [https://proceedings.neurips.cc/paper\\_files/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf)
- [5] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2023. A Survey on Evaluation of Large Language Models. *arXiv preprint arXiv:2307.03109* (2023).
- [6] Lijia Chen, Pingping Chen, and Zhijian Lin. 2020. Artificial Intelligence in Education: A Review. *IEEE Access* 8 (2020), 75264–75278. <https://doi.org/10.1109/ACCESS.2020.2988510>
- [7] Maximillian Chen, Alexandros Papangelis, Chenyang Tao, Seokhwan Kim, Andy Rosenbaum, Yang Liu, Zhou Yu, and Dilek Hakkani-Tur. 2023. PLACES: Prompting Language Models for Social Conversation Synthesis. In *Findings of the Association for Computational Linguistics: EACL 2023*. Association for Computational Linguistics, Dubrovnik, Croatia, 844–868. <https://aclanthology.org/2023.findings-eacl.63>
- [8] Maximillian Chen, Alexandros Papangelis, Chenyang Tao, Andy Rosenbaum, Seokhwan Kim, Yang Liu, Zhou Yu, and Dilek Hakkani-Tur. 2022. Weakly supervised data augmentation through prompting for dialogue understanding. In *NeurIPS 2022 Workshop on SyntheticData4ML*. <https://www.amazon.science/publications/weakly-supervised-data-augmentation-through-prompting-for-dialogue-understanding>
- [9] Maximillian Chen, Xiao Yu, Weiyan Shi, Urvi Awasthi, and Zhou Yu. 2023. Controllable Mixed-Initiative Dialogue Generation through Prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Toronto, Canada, 951–966. <https://doi.org/10.18653/v1/2023.acl-short.82>
- [10] Xieling Chen, Haoran Xie, Di Zou, and Gwo-Jen Hwang. 2020. Application and theory gaps during the rise of Artificial Intelligence in Education. *Computers and Education: Artificial Intelligence* 1 (2020), 100002. <https://doi.org/10.1016/j.caeai.2020.100002>
- [11] Okonkwo Chinedu and Abejide Ade-Ibijola. 2021. Python-Bot: A Chatbot for Teaching Python Programming. *Engineering Letters* 29 (02 2021), 25–34.
- [12] Samuel Cunningham-Nelson, Wageeh Boles, Luke Trouton, and Emily Margerison. 2019. *A Review of Chatbots in Education: Practical Steps Forward*. Engineers Australia, Brisbane, Queensland, 299–306. <https://search.informit.org/doi/10.3316/informit.068364390172788>

- [13] St. John Fisher. 2016. The Use of Technology to Support Vocabulary Development of English Language Learners. <https://api.semanticscholar.org/CorpusID:20174813>
- [14] Francisco-Javier Hinojo-Lucena, Inmaculada Aznar-Díaz, María-Pilar Cáceres-Reche, and José-María Romero-Rodríguez. 2019. Artificial Intelligence in Higher Education: A Bibliometric Study on its Impact in the Scientific Literature. *Education Sciences* 9, 1 (2019). <https://doi.org/10.3390/educsci9010051>
- [15] Caroline M.L. Ho, Mark Evan Nelson, and Wolfgang Müller-Wittig. 2011. Design and implementation of a student-generated virtual museum in a language curriculum to enhance collaborative multimodal meaning-making. *Computers & Education* 57, 1 (2011), 1083–1097. <https://doi.org/10.1016/j.compedu.2010.12.003>
- [16] Christopher L Holden and Julie M Sykes. 2011. Leveraging mobile games for place-based language learning. *International Journal of Game-Based Learning (IJGBL)* 1, 2 (2011), 1–18.
- [17] Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and Yejin Choi. 2022. SODA: Million-scale Dialogue Distillation with Social Commonsense Contextualization. *ArXiv abs/2212.10465* (2022).
- [18] Hyunwoo Kim, Youngjae Yu, Liwei Jiang, Ximing Lu, Daniel Khashabi, Gunhee Kim, Yejin Choi, and Maarten Sap. 2022. ProsocialDialog: A Prosocial Backbone for Conversational Agents. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 4005–4029. <https://doi.org/10.18653/v1/2022.emnlp-main.267>
- [19] Mojtaba Komeili, Kurt Shuster, and Jason Weston. 2022. Internet-Augmented Dialogue Generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, 8460–8478. <https://doi.org/10.18653/v1/2022.acl-long.579>
- [20] Stephen Krashen. 1982. Principles and practice in second language acquisition. (1982).
- [21] Mohammad Amin Kuhail, Nazik Alturki, Salwa Alramlawi, and Kholood Alhejori. 2023. Interacting with educational chatbots: A systematic review. *Education and Information Technologies* 28, 1 (2023), 973–1018.
- [22] Andrew S. Lan, Andrew E. Waters, Christoph Studer, and Richard G. Baraniuk. 2014. Sparse factor analysis for learning and content analytics. *Journal of Machine Learning Research* 15 (2014), 1959 – 2008. <https://www.scopus.com/inward/record.uri?eid=2-s2.0-84904196866&partnerID=40&md5=b3169cf7b6c6c25ca42fb8818b2b121e> Cited by: 96.
- [23] Yu Li, Chun-Yen Chen, Dian Yu, Sam Davidson, Ryan Hou, Xun Yuan, Yinghua Tan, Derek Pham, and Zhou Yu. 2022. Using Chatbots to Teach Languages. In *Proceedings of the Ninth ACM Conference on Learning @ Scale* (New York City, NY, USA) (*L@S '22*). Association for Computing Machinery, New York, NY, USA, 451–455. <https://doi.org/10.1145/3491140.3528329>
- [24] V Murphy, Henriette Arndt, Jessica Briggs Baffoe-Djan, Hamish Chalmers, Ernesto Macaro, Heath Rose, Robert Vanderplank, and Robert Woore. 2020. Foreign language learning and its impact on wider academic outcomes: A rapid evidence assessment. (2020).
- [25] Chinedu Wilfred Okonkwo and Abejide Ade-Ibijola. 2021. Chatbots applications in education: A systematic review. *Computers and Education: Artificial Intelligence* 2 (2021), 100033. <https://doi.org/10.1016/j.caeai.2021.100033>
- [26] OpenAI. 2023. GPT-4 Technical Report. *arXiv:2303.08774* [cs.CL]
- [27] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. 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 Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *arXiv:2203.02155* [cs.CL]
- [28] Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116* (2023). *arXiv:2306.01116* <https://arxiv.org/abs/2306.01116>
- [29] Kun Qian, Ryan Shea, Yu Li, Luke Fryer, and Zhou Yu. 2023. User Adaptive Language Learning Chatbots with a Curriculum.
- [30] Adam Roberts, Colin Raffel, Katherine Lee, Michael Matena, Noam Shazeer, Peter J. Liu, Sharan Narang, Wei Li, and Yanqi Zhou. 2019. *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*. Technical Report. Google.
- [31] Fátima Rodrigues and Paulo Oliveira. 2014. A system for formative assessment and monitoring of students' progress. *Computers & Education* 76 (2014), 30–41. <https://doi.org/10.1016/j.compedu.2014.03.001>
- [32] Mónica Rodríguez-Castro. 2018. An integrated curricular design for computer-assisted translation tools: developing technical expertise. *The Interpreter and Translator Trainer* 12 (07 2018), 1–20. <https://doi.org/10.1080/1750399X.2018.1502007>
- [33] Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, et al. 2022. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage. *arXiv preprint arXiv:2208.03188* (2022).
- [34] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv:2302.13971* [cs.CL]
- [35] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina



- Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288 [cs.CL]
- [36] Davis University of California. 2019. Gunrock 2.0: A user adaptive social conversational system. In *Alexa Prize SocialBot Grand Challenge 3 Proceedings*. <https://www.amazon.science/alexa-prize/proceedings/gunrock-2-0-a-user-adaptive-social-conversational-system>
- [37] Deeksha Varshney, Aizan Zafar, Niranshu Kumar Behera, and Asif Ekbal. 2023. Knowledge grounded medical dialogue generation using augmented graphs. *Scientific Reports* 13, 1 (2023), 3310.
- [38] Marco Antonio Rodrigues Vasconcelos and Renato P. dos Santos. 2023. Enhancing STEM learning with ChatGPT and Bing Chat as objects to think with: A case study. *Eurasia Journal of Mathematics, Science and Technology Education* 19, 7 (jul 2023), em2296. <https://doi.org/10.29333/ejmste/13313>
- [39] Ben Wang. 2021. Mesh-Transformer-JAX: Model-Parallel Implementation of Transformer Language Model with JAX. <https://github.com/kingoflolz/mesh-transformer-jax>.
- [40] Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is ChatGPT a Good NLG Evaluator? A Preliminary Study. arXiv:2303.04048 [cs.CL]
- [41] Ke Zhang and Ayse Begum Aslan. 2021. AI technologies for education: Recent research & future directions. *Computers and Education: Artificial Intelligence* 2 (2021), 100025. <https://doi.org/10.1016/j.caeai.2021.100025>
- [42] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: Open Pre-trained Transformer Language Models. arXiv:2205.01068 [cs.CL]
- [43] Chujie Zheng, Sahand Sabour, Jiaxin Wen, Zheng Zhang, and Minlie Huang. 2023. AugESC: Dialogue Augmentation with Large Language Models for Emotional Support Conversation. In *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics, Toronto, Canada, 1552–1568. <https://doi.org/10.18653/v1/2023.findings-acl.99>
- [44] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. arXiv:2306.05685 [cs.CL]

## A DATA AUGMENTATION EXAMPLES

### A.1 Example of Topic Augmentation Using ChatGPT

An example is presented below to demonstrate how ChatGPT is used to generate related subtopics from a main topic in the textbook.

- Primary Topic:  
The true value of education
- Prompt to ChatGPT:  
Given an input topic, generate a list of 10 closely related topics that could be explored further.  
Input topic: The true value of education
- Generated Sub-topics:  
“The importance of education in personal and professional development”,  
“The impact of education on social and economic mobility”,  
“The relationship between education and individual well-being”,  
“The role of education in promoting social justice and equity”,  
“The benefits of a well-rounded education that includes arts, humanities, and social sciences”,  
“The value of education in fostering critical thinking and problem-solving skills”,  
“The potential of education in promoting innovation and entrepreneurship”,  
“The relationship between education and global competitiveness”,  
“The significance of lifelong learning in today’s rapidly changing world”,

“The need for education reform to address societal challenges and prepare students for the future”

This example demonstrates the capability of ChatGPT to take a given topic and generate a variety of associated sub-topics that expand the breadth of our conversational database.

## A.2 Example of Synthesizing Personas of Participants Using ChatGPT

An example is presented below to demonstrate how ChatGPT is used to generate personas of participants in a conversation.

- Prompt to ChatGPT:

Please provide me with one individual Person 1 with different backgrounds, including information about their demographic, socio-economic status, culture, MBTI personality type, and personal experiences, no need to show names. Then provide me with one individual Person 2 who is a college student originating from China but with different information.

- Generated Personas:

Person 1:

Gender: Male

Demographic: African American

Socio-economic status: Working class

Culture: Baptist

MBTI personality type: ENFP

Personal experience: He grew up in a low-income neighborhood, and as the oldest child, he felt responsible for taking care of his siblings. He struggled with academics in high school but managed to graduate. He is now working as a bartender and uses his outgoing personality to make friends with his customers.

Person 2:

Gender: Female

Demographic: Chinese

Socio-economic status: Upper middle class

Culture: Confucianism

MBTI personality type: INTP

Personal experience: She comes from a wealthy family and has had access to quality education throughout her life. She is currently studying computer science in college and hopes to start her own tech company after graduation. Growing up in a traditional Chinese household, she feels pressure from her parents to succeed academically and make the family proud. She struggles with balancing her ambitious career goals with her desire for independence and freedom.

This example demonstrates that ChatGPT can generate personas in a fixed format and follows our instructions that Person 2 is a college student with other random personas.

### A.3 Example of Synthesizing a Dialogue Using ChatGPT

We demonstrate the use of ChatGPT in our approach by providing an example. We use the personas in Appendix A.2 and the topic augmented in Appendix A.1 to create a dialogue. The words used are taken from the vocabulary related to the topic “The true value of education” in the textbook.

- Prompt to ChatGPT:

Generate a single conversation between these two people as Person 1 and Person 2 about the topic “The importance of education in personal and professional development”.

Please take into account their distinct personalities and their backgrounds. Begin the conversation with Person 1.

Please include the following keywords in Person 1’s utterances: “due”, “get down to”, “get away with”, “slam dunk”, “the easy way out”, “hand out”, “supposedly”, “revenue”, “industrial”, “rustle”

Person 1 should guide the conversation by asking more questions

- Generated Dialogue:

Person 1: Hey, have you ever thought about the significance of education in personal and professional development? People sometimes believe that they can get away with not studying and still become successful, but that’s not always the case.

Person 2: Yes, I completely agree with you on that. Education paves the way for a successful career and personal growth. It provides the necessary skills and knowledge to make informed decisions and take on challenges in life.

Person 1: I see that you’re a computer science major. What made you choose this field?

Person 2: Well, I’ve always been interested in technology and innovation. I think computer science is an industry that is constantly growing and provides a lot of opportunities for revenue and growth.

Person 1: That’s interesting. Do you think attending college is a requirement for success in computer science, or do you think there are other routes to succeed?

Person 2: I think college is definitely a great way to learn about the industry and gain practical skills. However, there might be alternative ways of gaining knowledge and experience in the industry too. What do you think, Person 1?

Person 1: I believe education is important in any field, whether it’s through traditional schooling or hands-on experience. Sometimes it’s easier to take the easy way out and just hope someone will hand you an opportunity, but usually, you have to get down to work and rustle up some opportunities yourself. Sometimes people think there’s a magic formula to success, but there’s not a slam dunk for anyone. You need to put in the effort and invest

in yourself. What are your thoughts on that?

Person 2: I couldn't agree more. Education provides a solid foundation for professional development, but it's also important for personal growth. Knowing how to learn and adapt to change is essential in today's fast-paced and industrial world. Education is not just about getting a degree or a job, it's about being a lifelong learner and constantly upgrading yourself. What do you think about that?

Person 1: I think that's a smart way of seeing things! Education is supposed to be an investment in ourselves, not just for our future jobs or salaries. It gives us the tools to think critically, challenge ourselves, and grow in all areas of life. It's like learning a language - just like in school, you study grammar first before you can speak more fluently. It's the same with learning about other subjects. You're not going to jump into the field without basics. But once you have a foundation, you can build up from there.

This example illustrates how ChatGPT can generate conversations between Person 1 and Person 2 based on our instructions. We can observe that the dialogue revolves around the topic we specify, and most of the words we provide are used in the dialogue. Furthermore, both participants incorporate their individual experiences of their personas into the conversation.

## B IMPLEMENTATION DETAILS

To train a model for our application, we choose the 13B Vicuna model<sup>5</sup>. During the training phase, we carefully match each turn of our generated dialogues with the corresponding training turn in Vicuna format. As mentioned in Section 3.1.2, Person 1's persona represents the chatbot's side, while Person 2's persona represents the students'. Therefore, we use utterances from Person 1 as the system's responses and those from Person 2 as user requests throughout our training process. We train the Vicuna model for 3 epochs, beginning with a learning rate of  $2e-5$ . We use a batch size of 1 on each GPU and a gradient accumulation step of 16. We utilize 8 A100 GPUs and the training process takes three hours to complete.

## C BACKGROUND SURVEY AND QUESTIONNAIRE

### C.1 Background Survey

Table 2 shows the full background survey we used for recruiting participants. "College English 4" uses the "New College English" (3rd edition) textbook and is a mandatory course for student participants of our user study. CET-4 and CET-6 are standardized English proficiency exams for Chinese college students.

### C.2 Questionnaire

Table 3 presents the questionnaire we used to compare the quality of EduBot and ChatGPT from various aspects.

Table 2. Background Survey for User Study Participants

Number	Question
1	Student ID
2	WeChat ID
3	Gender
4	Age
5	Grade
6	Major
7	Duration of English Learning
8	Overall Grade for <i>College English 4</i>
9	CET-4 Total Score
10	CET-4 Examination Date
11	CET-6 Total Score
12	CET-6 Examination Date
13	Available Time Slots

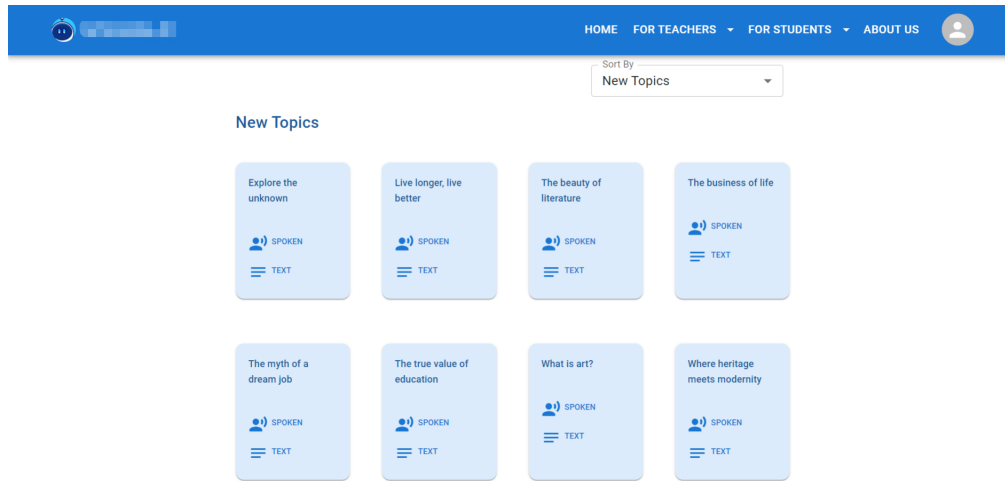


Fig. 10. User Interface for Selecting a Textbook Unit as the Conversation Topic

## D USER INTERFACE

We used the following user interface for both EduBot and ChatGPT. The user first selects a unit from the textbook (Figure 10) as the main topic of conversation, then proceeds to chat with the bot (Figure 11).

## E ANALYSIS OF PARTICIPANTS' ENGLISH PROFICIENCY LEVELS

In this section, we analyze the influence of participants' English proficiency levels on their perception of the two chatbots. We divided the participants into the following three groups according to their overall grade for the course "College English 4": Group A consists of 8 students with scores between 2.1 and 3.6, Group B of 10 students with scores between 3.9 and 4.5, and Group C of 6 students with scores between 4.8 and 5.0. We reached the following conclusions.

<sup>5</sup><https://lmsys.org/blog/2023-03-30-vicuna/>

Table 3. Questionnaire

Section	Number	Question
Participant Information	1	Student ID
Dialogue Summarization	2	Please summarize the main content of your first conversation with chatbot A.
	3	Please summarize the main content of your second conversation with chatbot A.
	4	Please summarize the main content of your first conversation with chatbot B.
	5	Please summarize the main content of your second conversation with chatbot B.
Consistency with Curriculum	6-1	The main topics of my conversations with the chatbot were closely related to what I learned in English class.
	6-2	The chatbot brought up anecdotes, examples, questions, etc., related to what I learned in English class.
	6-3	The chatbot mentioned topics and content that were not directly covered in the textbook and course.
English Proficiency Level	7-1	During our conversations, the chatbot mentioned some vocabulary words that I learned in my English course.
	7-2	The chatbot used many vocabulary words that I didn't understand.
	7-3	I didn't find the conversations too easy to be helpful.
Role Identification	8-1	During conversations, I felt that the chatbot recognizes that I am a Chinese college student.
	8-2	During the two conversations with the chatbot, I felt like I was talking with two different people.
Conversation Language Quality	9-1	The utterances provided by the chatbot were coherent and fluent.
	9-2	The chatbot's responses were concise and accurate.
	9-3	Unlike in real everyday conversations, the chatbot's responses were long and redundant at times.
	9-4	Interactions with the bot were similar to natural, realistic conversations and not overly formal.
Conversation Content Quality	10-1	The chatbot acknowledged what I said and provided reasonable responses.
	10-2	The chatbot provided unique and personal perspectives regarding the selected topic.
	10-3	The chatbot used personal experiences to support its opinions.
	10-4	The chatbot actively raised questions to guide the course of the conversation.
	10-5	The chatbot didn't output offensive or hurtful responses.
General Usefulness	11-1	I would find it useful to use the chatbot to review what I learned in class.
	11-2	I would recommend the chatbot to other students.
	11-3	I believe that continuing to use the chatbot will help me improve my English conversation skills.

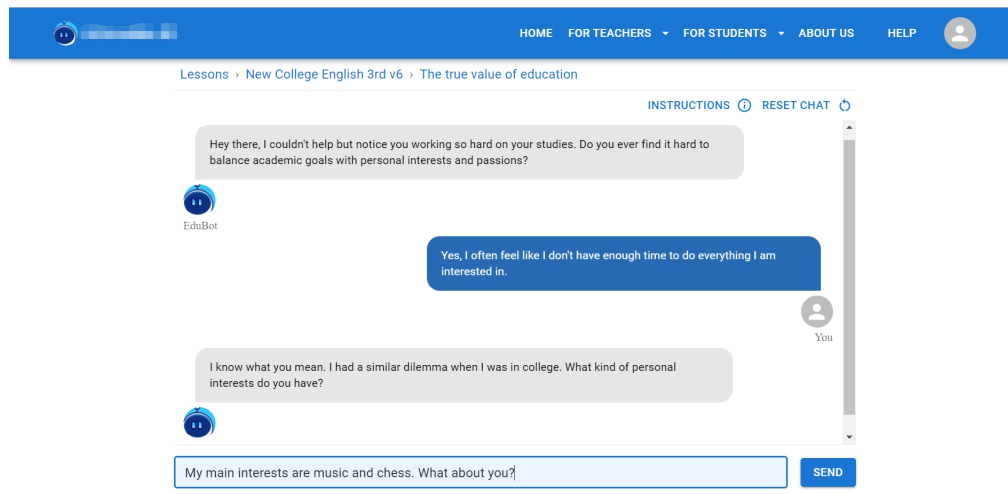


Fig. 11. User Interface for Conversing with the Chatbots



Fig. 12. Participants with lower English proficiency levels found it more difficult to distinguish between the two chatbots.

*E.0.1 Participants with lower English proficiency levels found it more difficult to distinguish between the two chatbots.* We observed that students in Group A were more likely to believe that the two chatbots performed the same over multiple questions. In addition, their responses were more often evenly split between the two chatbots. To verify, we calculated the following two statistics separately for each group of students: the average win rate of the “same” option over all questions and the average difference between win rates of “EduBot” and “ChatGPT” over all questions. The results are shown in Figure 12. We believe this is because it was harder for students in Group A to understand the chatbots and fully engage in the conversation.

*E.0.2 Participants with high English proficiency levels were more likely to prefer EduBot.* In Figure 13, we present the three groups’ win rate results for the final section of the questionnaire. For the criteria “11-2 I would recommend the chatbot to other students” and “11-3 I believe that continuing to use the chatbot will help me improve my English conversation skills”, all participants in Group C chose either “EduBot” or “Same”. For “11-1 I would find it useful to use the chatbot to review what I learned in class”, results from Group C were in line with results from all the participants

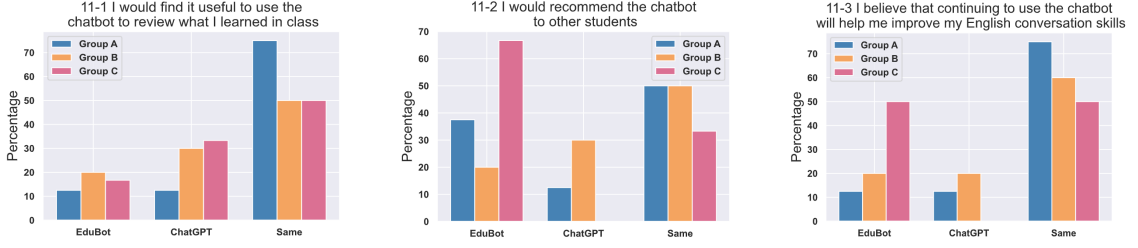


Fig. 13. Participants with high English proficiency levels were more likely to prefer EduBot.

combined, with ChatGPT slightly outperforming EduBot. We believe that students in Group C more strongly preferred EduBot as a conversational training tool because they were more inclined to actively engage in conversations and provide their own thoughts instead of passively responding to the chatbot’s utterances. This caused EduBot’s advantages of providing natural responses and guiding the conversation by asking questions to be underscored in Group C’s results.

## F EDUBOT-STUDENT CONVERSATION EXAMPLES

We present three conversations between our EduBot and a student from our user study in Figures 14, 15, and 16.

## G ANALYSIS OF USER STUDY CONVERSATIONS

We extracted all conversation histories from our user study. In the following section, we analyze the utterance lengths and coverage of target vocabulary words in the user study conversations.

### G.1 Utterance Lengths

As shown in Figure 19, we observe that in our user studies, ChatGPT produced longer outputs compared with EduBot. ChatGPT’s utterances were on average approximately 10 words longer than EduBot’s. In addition, ChatGPT occasionally produced outputs that were longer than 60 words, which rarely occurs in natural, daily conversations.

Furthermore, Figures 17 and 18 demonstrate that user study participants generally provided longer responses when conversing with EduBot compared to ChatGPT. This indicates that EduBot’s more interactive and realistic conversation style better engages the users and guides them to practice their own conversation skills.

### G.2 Target Vocabulary Words

We also assess if EduBot can incorporate words from the target vocabulary. As shown in Figure 20, on average, conversations with EduBot included 5.55 words from the target vocabulary, while conversations with ChatGPT only included 0.62. This demonstrates that EduBot, which was further refined using curriculum-aligned data, is better suited to the user’s curriculum and English level.



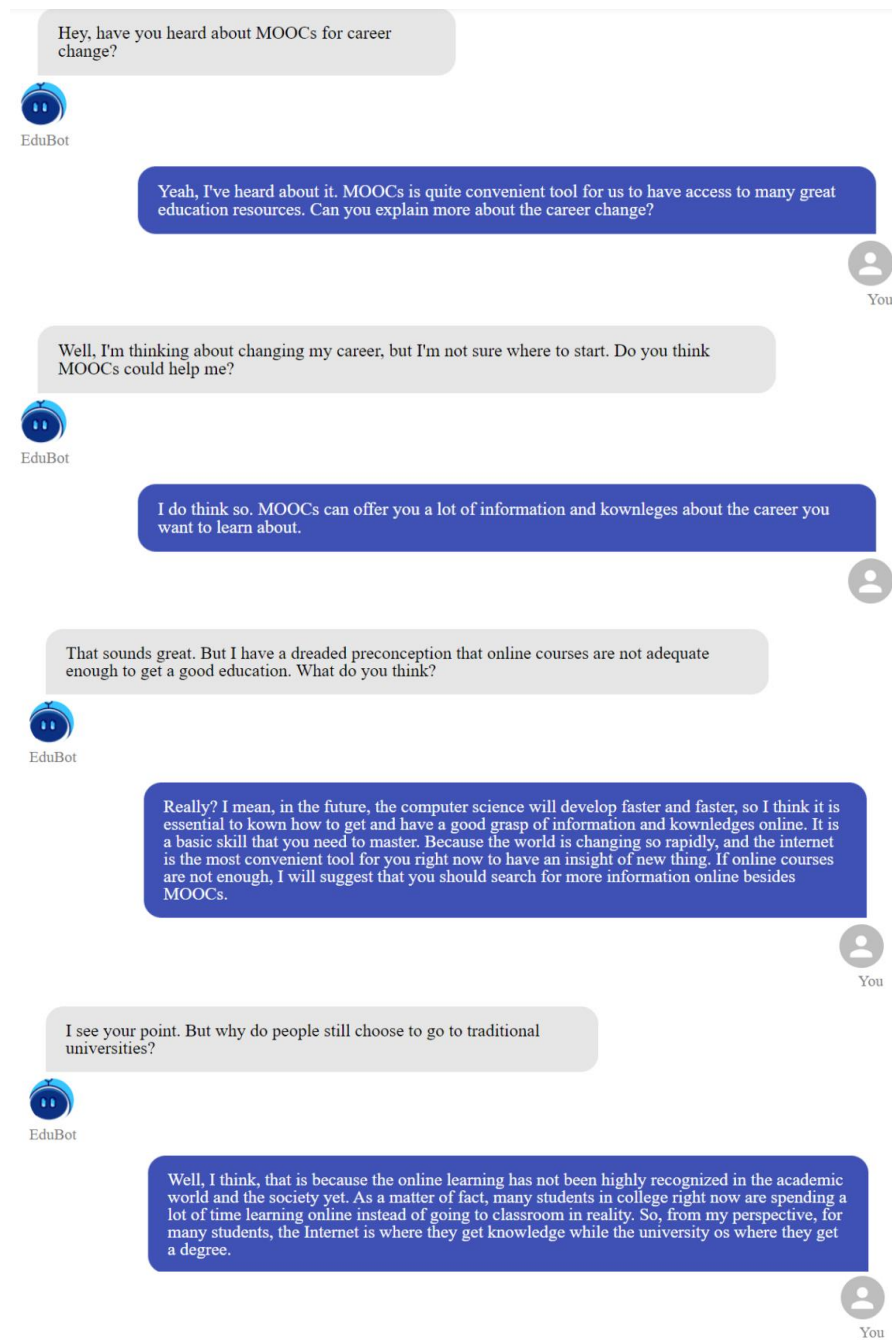


Fig. 14. Example conversation from our user study (1)

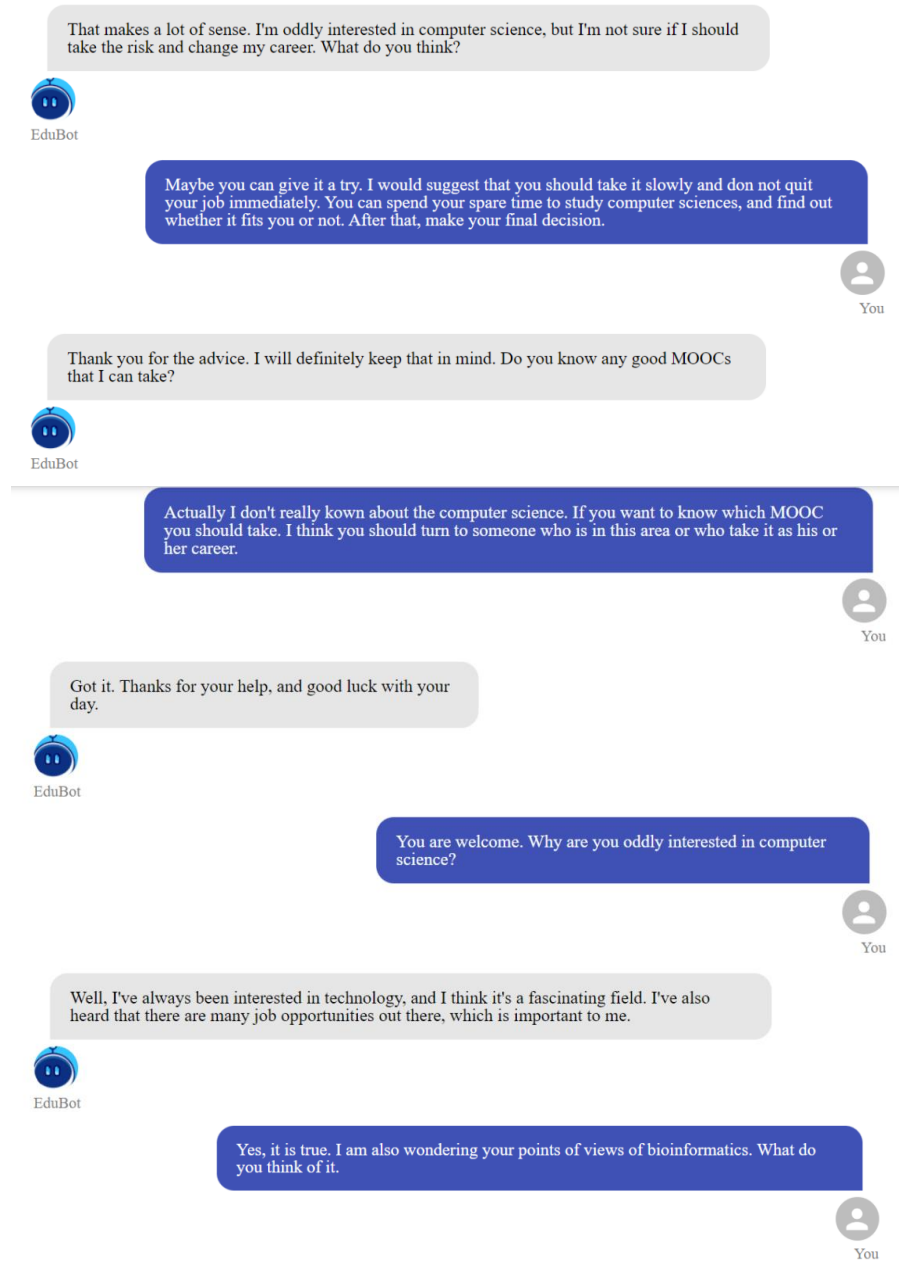


Fig. 15. Example conversation from our user study (2))

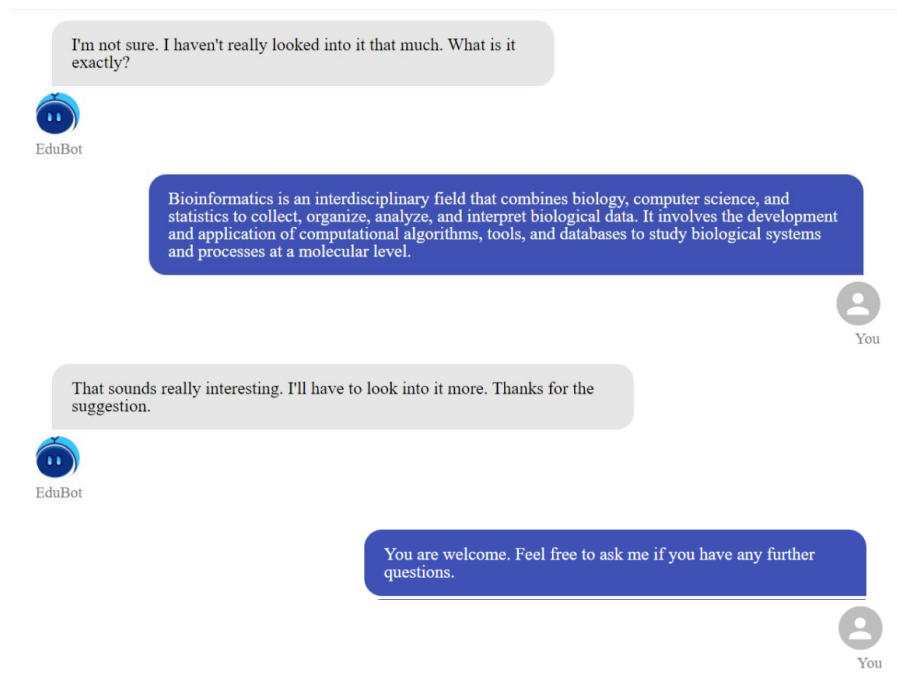


Fig. 16. Example conversation from our user study (3)

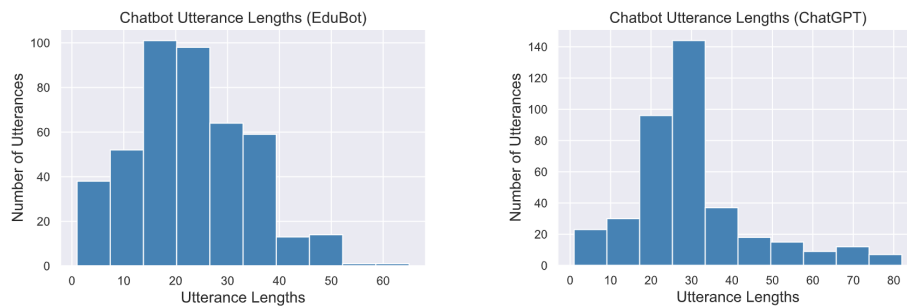


Fig. 17. Lengths of chatbot utterances in the user study

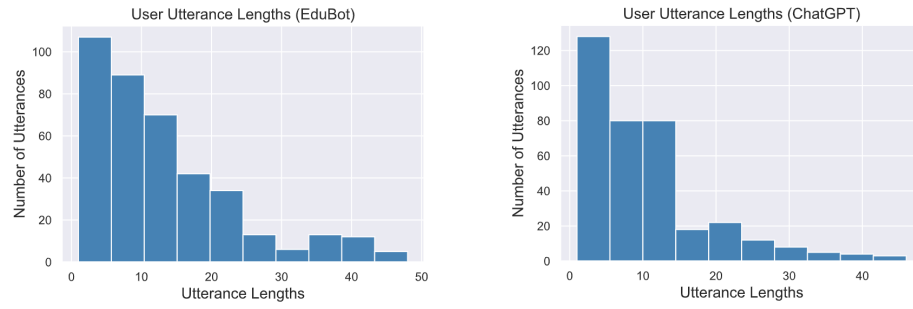


Fig. 18. Lengths of user utterances in the user study

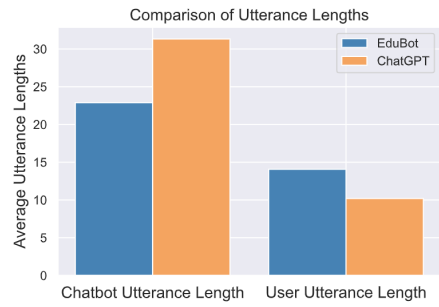


Fig. 19. Comparison of utterance lengths in EduBot and ChatGPT conversations in the user study

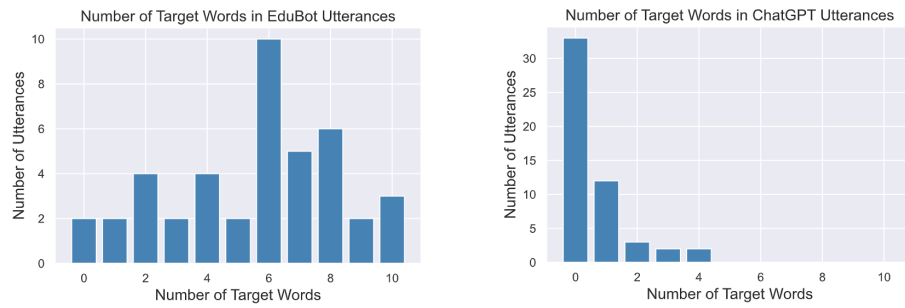


Fig. 20. Coverage of target words in user study conversations