Department of Electrical & Computer Engineering

ECE 1786: Creative Applications of Natural Language Processing

Professor. Jonathan Rose

LangoBot: Final Report

Xinyu Liu – 1004723640
Lanzhi Xiao – 1004569362

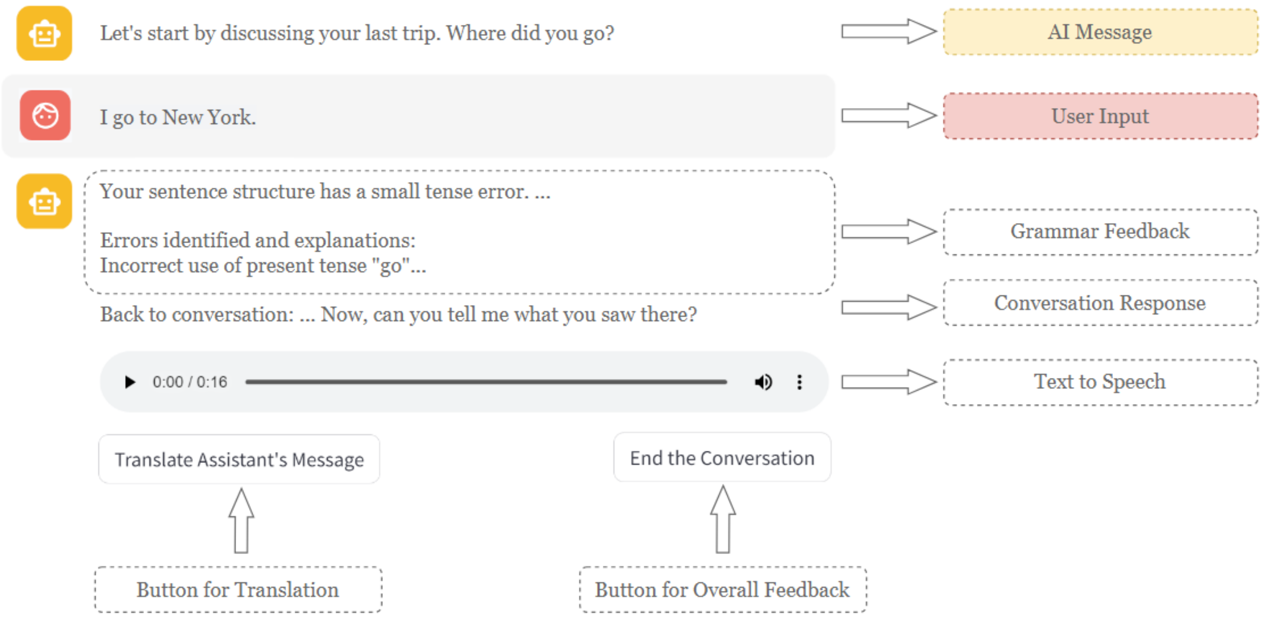
Word Count: 1998 (Exclude all figures and captions but include table 2 and 5)
Penalty: 0%

0. Permissions

Member	Post Video?	Post Report?	Post Code?
Xinyu Liu	No	Yes	Yes
Lanzhi Xiao	No	Yes	Yes

1. Introduction

Traditional English classes often lack accessibility and personalization, and many learners may not feel comfortable speaking with a human instructor. Addressing these challenges, our project introduces “LangoBot”, an English learning chatbot that utilizes a large language model (LLM). This solution aims to improve users' English skills through AI-driven, interactive conversations, customized to each user's proficiency level. In our application, the AI acts as a teacher, leading discussions on users’ topics of interest or guiding them through English practice sessions. Users benefit from instant grammar feedback, translation support, and comprehensive advice on their English skills.



Let's start by discussing your last trip. Where did you go?

I go to New York.

Your sentence structure has a small tense error. ...

Errors identified and explanations:
Incorrect use of present tense "go"...

Back to conversation: ... Now, can you tell me what you saw there?

Figure 1. General Illustration of LangoBot

2. Background

Recent advancements in Natural Language Processing (NLP) have paved the way for innovative approaches in language learning. One key research has focused on using open domain chatbots for second language acquisition [1] This research involved developing methods to customize AI responses to learners' proficiency levels and demonstrated AI-driven chatbots' potential for personalized learning experiences. The evaluation of these chatbots' message quality confirmed their capability to facilitate accurate and relevant language learning interactions.

Another study dives into the broader applications of Generative Artificial Intelligence (GAI) in the academic industry. LLMs like GPT-4, have been adopted by platforms like Duolingo for language learning [2]. Duolingo's features, like “Explain My Answer” and “Roleplay” demonstrated GPT-4's utility in language education. 'LangoBot' builds upon these advancements, integrating conversational responses tailored to users’ proficiency levels with instant grammar feedback to create a more interactive and adaptive language learning tool.

3. Architecture

Our project utilizes Streamlit Framework for its user interface and GPT-4 API for interaction logic. The system is driven by four main prompts, as illustrated in Figure 2, and summarized below. The full set of prompts can be found in the “*prompts.py”* file from our GitHub repository.

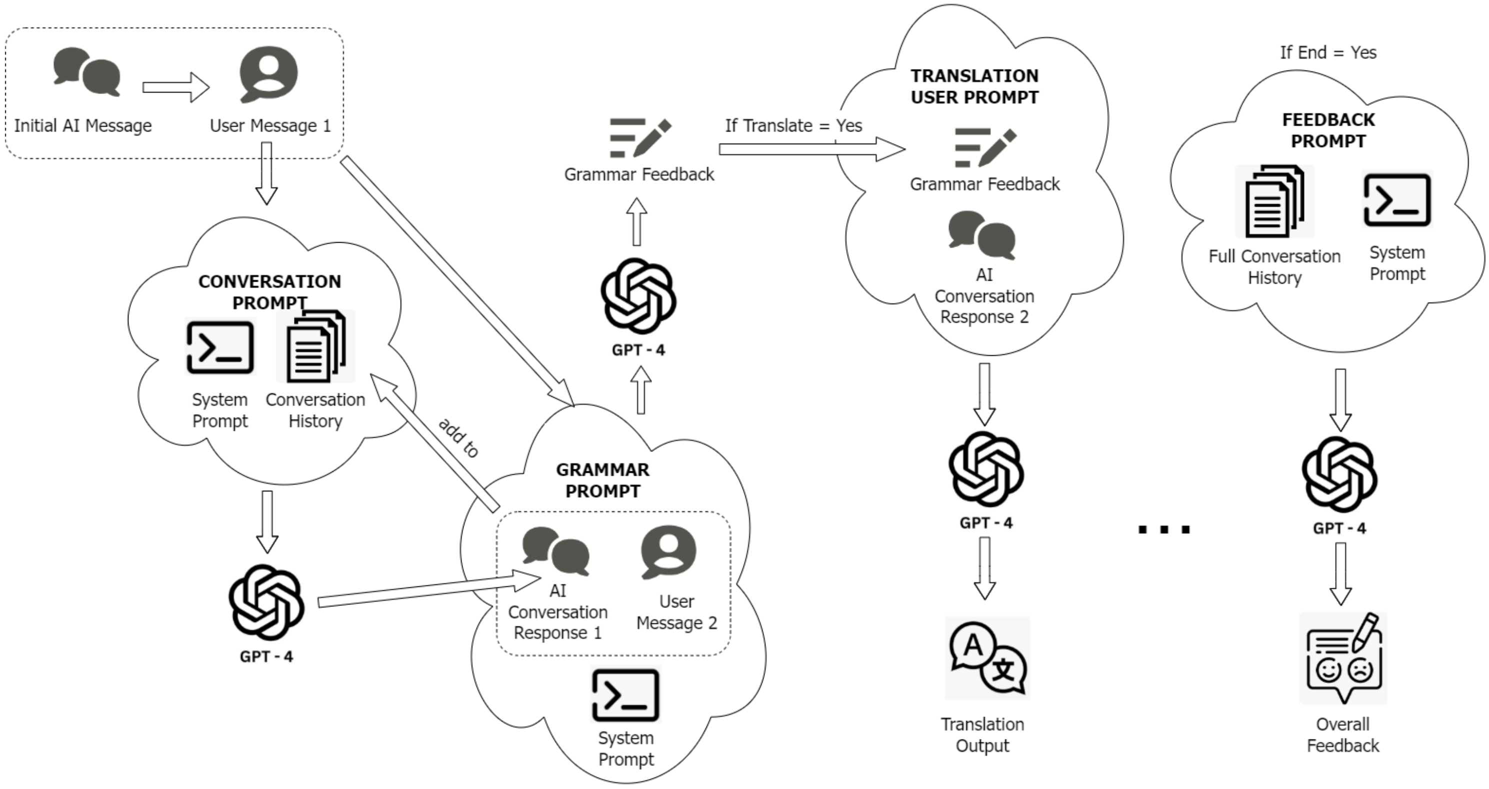


Figure 2. Prompt Interactions in LangoBot

Conversation Prompt: Begin by inquiring about users’ proficiency and interests to tailor an educational dialogue. In cases of uncertainty, suggest random topics or practice sessions. Engage users through varied interactions, including role-play scenarios.

The history of all AI-user messages is saved and repeatedly sent to GPT-4 with this prompt to generate context-aware AI conversation responses.

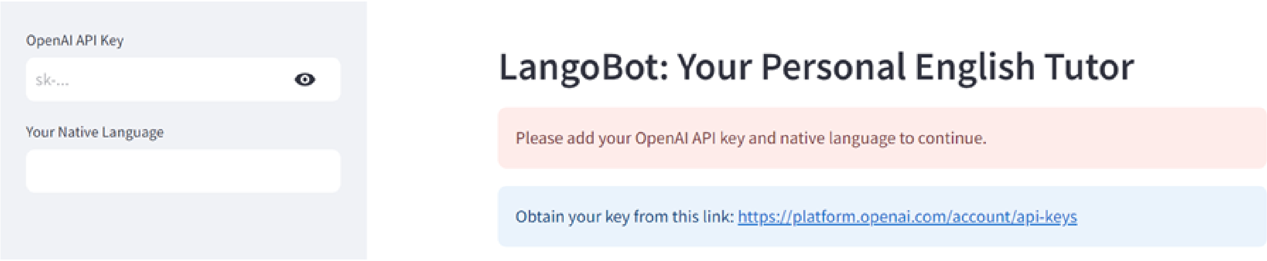
Grammar Prompt (Few-Shot): Analyze an AI-user message pair to identify grammatical errors in the user's response, provide grammar rule explanations and usage examples, determine response relevance to AI's message, and suggest an improved version of the response.

Translation Prompt: Translate a message from an English instructor into the user's native language.

Evaluation Prompt: Review conversation history to assess and provide feedback on users' overall English proficiency, focusing on vocabulary and grammar usage, reading comprehension, writing skills, and potential improvements.

In a practical scenario, Luna, a beginner learning English for visiting Canada, engages with LangoBot. She begins by entering her API key and native language (see Figure 3). The Conversation Prompt then leads Luna to specify her proficiency as a "Beginner" and her interest in travel (see Figure 4). LangoBot customizes the dialogue accordingly, focusing on travel-related topics and practice sessions involving travel phrases suitable for Luna's beginner level.

Throughout the interaction, the Grammar Prompt reviews Luna's responses, providing necessary corrections and explanations. Each AI response is also available in an audio format, enabling Luna to hear the AI's message for better pronunciation understanding. Luna can use the "Translate" button for additional translation support in her native language; one translated output is shown in Figure 5. At the end of her session, she can click the "End" button to receive comprehensive feedback on various language aspects, helping her track her learning progress. An example interaction can be viewed in Figure 1.



OpenAI API Key

sk-...

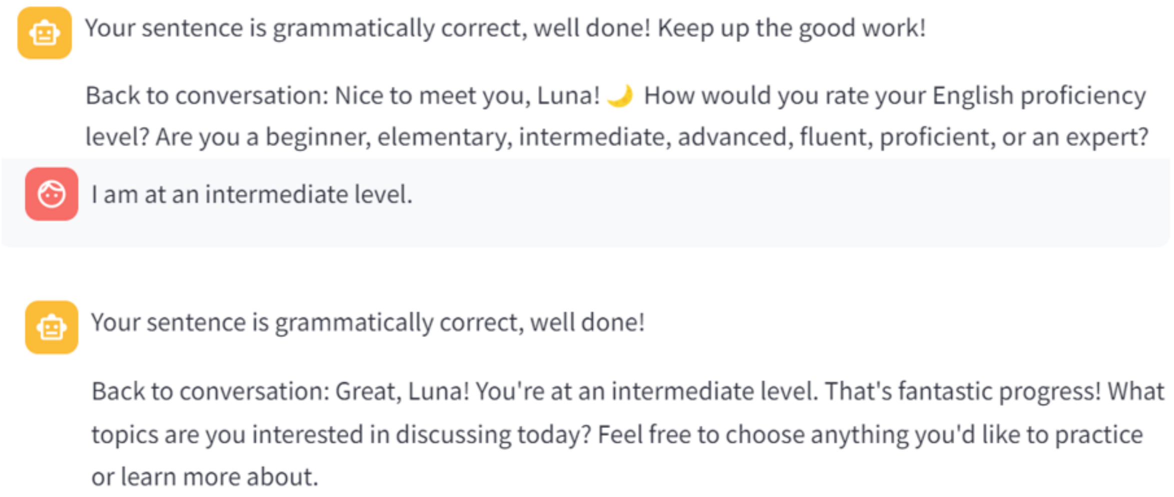
Your Native Language

LangoBot: Your Personal English Tutor

Please add your OpenAI API key and native language to continue.

Obtain your key from this link: https://platform.openai.com/account/api-keys

Figure 3. Initial Setup Screen of LangoBot



Your sentence is grammatically correct, well done! Keep up the good work!

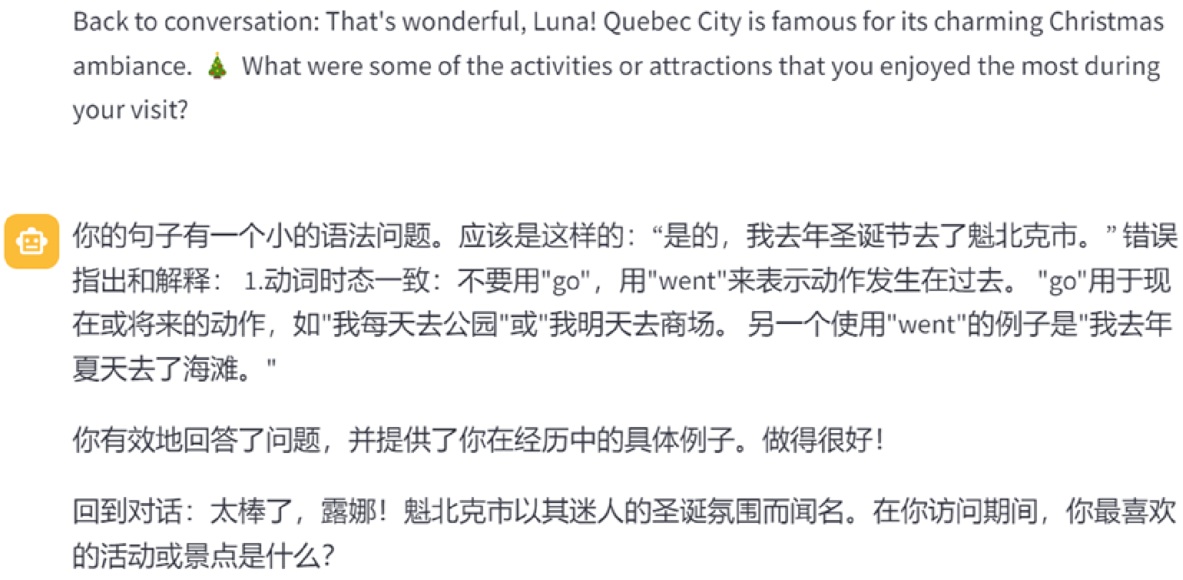
Back to conversation: Nice to meet you, Luna! 🌙 How would you rate your English proficiency level? Are you a beginner, elementary, intermediate, advanced, fluent, proficient, or an expert?

I am at an intermediate level.

Your sentence is grammatically correct, well done!

Back to conversation: Great, Luna! You're at an intermediate level. That's fantastic progress! What topics are you interested in discussing today? Feel free to choose anything you'd like to practice or learn more about.

Figure 4. Example of an Information Gathering Conversation



Back to conversation: That's wonderful, Luna! Quebec City is famous for its charming Christmas ambiance. 🎄 What were some of the activities or attractions that you enjoyed the most during your visit?

你的句子有一个小的语法问题。应该是这样的：“是的，我去年圣诞节去了魁北克市。”错误指出和解释： 1.动词时态一致：不要用"go"，用"went"来表示动作发生在过去。 "go"用于现在或将来的动作，如"我每天去公园"或"我明天去商场。 另一个使用"went"的例子是"我去年夏天去了海滩。"

你有效地回答了问题，并提供了你在经历中的具体例子。做得很好！

回到对话：太棒了，露娜！魁北克市以其迷人的圣诞氛围而闻名。在你访问期间，你最喜欢的活动或景点是什么?

Figure 5. Example of a Translated Message

4. Data and Success Evaluation

To comprehensively evaluate our system, three test datasets were generated to specifically target the assessment of key aspects including grammar correction, translation, and overall feedback performance. The evaluation metrics are detailed in the results section.

Grammar Dataset

80 message-response pairs were generated to assess the grammar correction capabilities. This involved writing additional prompts instructing GPT to respond as a user to the AI teacher in our system— creating an equal mix of error-free and grammatically incorrect responses. We reviewed the responses to ensure that the error-free ones were indeed without errors. Subsequently, we manually annotated each response with the count and types of errors, if any. An example is provided below:

Message: Great response! Next, let's talk about hobbies. Can you tell me what you like to do in your free time and why you enjoy doing it?

Response: In my free time, I like to read books. Reading books is make me feel relax.

Numbers of Mistakes: 2

Grammatical Errors: Agreement: "is make" should be "makes." Verb Form: "relax" should be "relaxed."

We then input each message-response pair into the Grammar Prompt to obtain the corresponding grammar feedback, thereby completing our grammar dataset.

Translation Dataset

Our application's translation function covers both AI conversation responses and grammar feedback. To assess this, we compiled a dataset of 160 examples by extracting these two elements—Message and Grammar Feedback—from the 80 grammar examples. Each element was translated into four languages: Chinese, German, French, and Spanish, using both our Translation Prompt and the DeepL Translation API. This resulted in a dataset where each example includes nine columns.

Feedback Dataset

To assess the overall feedback provided by our system, we created AI-user dialogues like those in the grammar dataset. We directed GPT to mimic users at various proficiency levels, occasionally introducing grammar errors or irrelevant responses. A total of 21 conversations were generated and subsequently processed by the Evaluation Prompt to obtain the overall feedback. Thus, this dataset is composed of two columns, conversation history and the corresponding feedback.

5. Results and Discussion

In this section, we will detail the evaluation metrics, present and discuss the quantitative and qualitative results for each functionality of our system separately.

Grammar Correction

Quantitative results

Our methods for evaluating grammar correction involved a manual comparison of the feedback generated by our system against predefined labels. The effectiveness of error identification was quantified using binary accuracy, precision, and recall. A binary accuracy score of 1 was assigned when both the count and types of errors identified by the system matched the labels. Partial accuracy was calculated as the number of correctly identified errors divided by the greater of the labeled error count or the number of errors identified by the system. This procedure is illustrated in Figure 6, and the results are shown in Table 1. These metrics are appropriate as they provide a clear, numerical assessment of the system's ability to correctly identify grammatical errors.

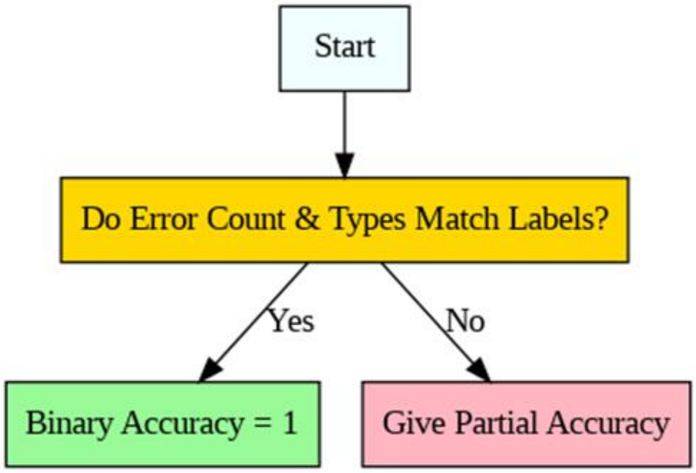


Figure 6. Quantitative Evaluation Procedure for Grammar Feedback

Metric	Binary Accuracy	Precision	Recall	Partial Accuracy
Value	0.9625	1.0	0.925	0.9812

Table 1. Grammar Feedback Performance Results

Qualitative results

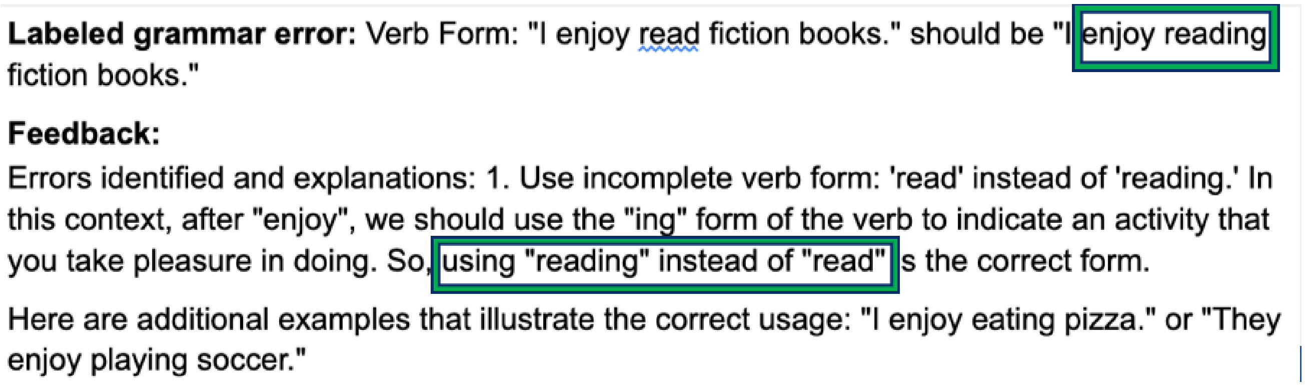
Beyond quantitative metrics, we introduced a custom qualitative metric to evaluate the grammar rule explanations. As evaluators, we assessed each explanation based on its clarity, accuracy, and information richness (IR), scoring each criterion from 0 to 2. The effectiveness of each error explanation was given a score out of 8, determined by multiplying accuracy with the sum of IR and clarity. We designed this scoring system to emphasize the importance of accuracy while also considering the overall usefulness of the explanations. The final normalized score for this qualitative assessment was **0.8691**. The detailed criteria used for this evaluation are outlined in Table 2.

Accuracy	(+0) Inaccurate: Completely incorrect grammar explanation.	(+1) Mostly Accurate: Partially correct explanation.	(+2) Accurate: Error-free explanation.
Clarity	(+0) Confusing: Unorganized, difficult to understand, lacking logical flow.	(+1) Understandable: Generally makes sense, may require additional review.	(+2) Clear: Well-structured, straightforward, easily comprehensible.
Information Richness	(+0) Poor: Lacks detailed explanations and examples.	(+1) Moderate: Includes useful explanations but missing specific examples.	(+2) Good: Detailed rule explanations with relevant examples.

Table 2. Grammar Explanation Evaluation Criteria

Figure 7 shows one grammar feedback that scores 8/8; it is clear and accurate, with detailed explanations and examples.

Figure 8 displays another example scores 3.67 out of 8. The second identified error scores 0 because it notes the misuse of 'the' but fails to accurately locate it. The third error is partially correct, it should point out missing 'a' before 'circle' instead of 'the,' resulting in a one-point deduction in IR and accuracy, leading to a score of 3. This example's overall score is calculated as (8+0+3)/3, totaling 3.67.



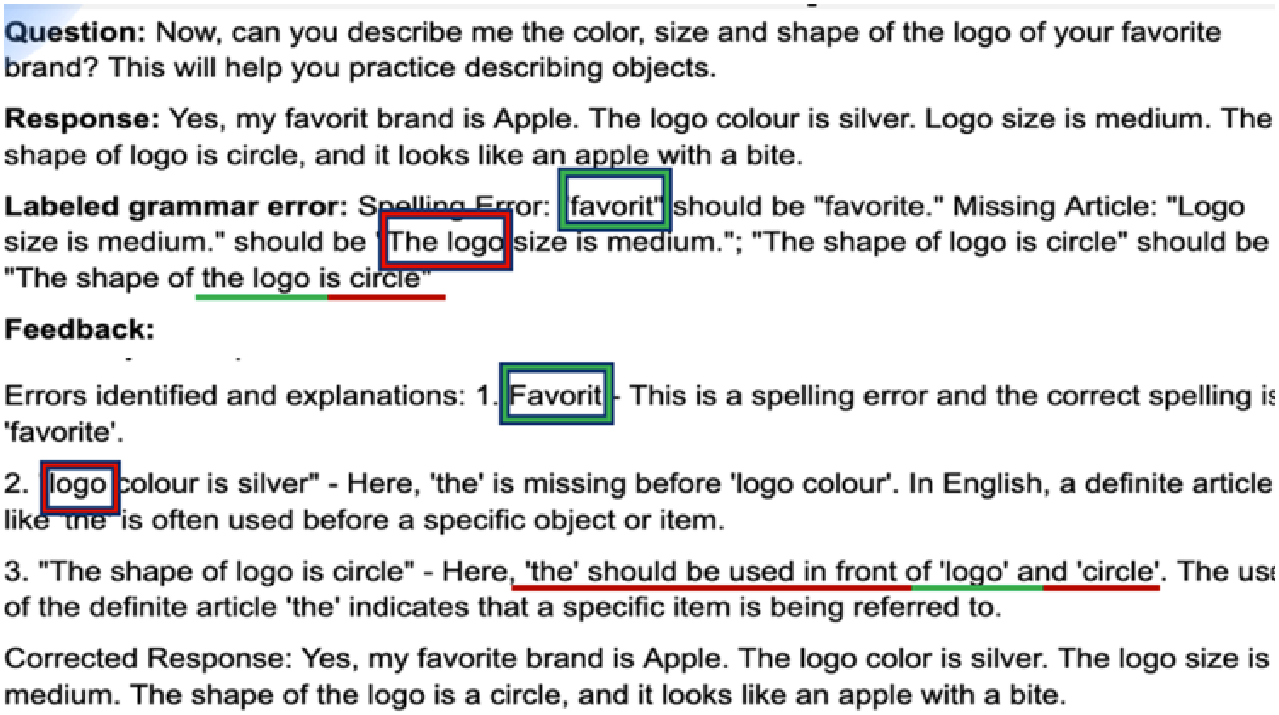
Labeled grammar error: Verb Form: "I enjoy read fiction books." should be "I enjoy reading fiction books."

Feedback:

Errors identified and explanations: 1. Use incomplete verb form: 'read' instead of 'reading.' In this context, after "enjoy", we should use the "ing" form of the verb to indicate an activity that you take pleasure in doing. So, using "reading" instead of "read" is the correct form.

Here are additional examples that illustrate the correct usage: "I enjoy eating pizza." or "They enjoy playing soccer."

Figure 7. An Example of Good Grammar Feedback



Question: Now, can you describe me the color, size and shape of the logo of your favorite brand? This will help you practice describing objects.

Response: Yes, my favorit brand is Apple. The logo colour is silver. Logo size is medium. The shape of logo is circle, and it looks like an apple with a bite.

Labeled grammar error: Spelling Error: "favorit" should be "favorite." Missing Article: "Logo size is medium." should be "The logo size is medium."; "The shape of logo is circle" should be "The shape of the logo is circle"

Feedback:

Errors identified and explanations: 1. Favorit - This is a spelling error and the correct spelling is 'favorite'.

2. "logo colour is silver" - Here, 'the' is missing before 'logo colour'. In English, a definite article like 'the' is often used before a specific object or item.

3. "The shape of logo is circle" - Here, 'the' should be used in front of 'logo' and 'circle'. The us of the definite article 'the' indicates that a specific item is being referred to.

Corrected Response: Yes, my favorite brand is Apple. The logo color is silver. The logo size is medium. The shape of the logo is a circle, and it looks like an apple with a bite.

Figure 8. An Example of Bad Grammar Feedback

Discussion:

The quantitative results are impressive, with a precision of 1 indicating the system reliably identifies errors without falsely flagging correct sentences. However, a recall of 0.925 suggests that the system occasionally misses errors, which needs improvement. Qualitatively, the grammar explanation score is decent. The system sometimes struggles to identify and explain errors in responses with too many mistakes, but it consistently provides corrected sentence versions, even in cases of inaccurate error identification.

Translation

Quantitative results and discussion

We utilized BLEU scores to evaluate our system's translation quality, which quantifies the similarity between LangoBot’s translations and those of DeepL. Scores nearing 0.6 indicate a very high degree of similarity.

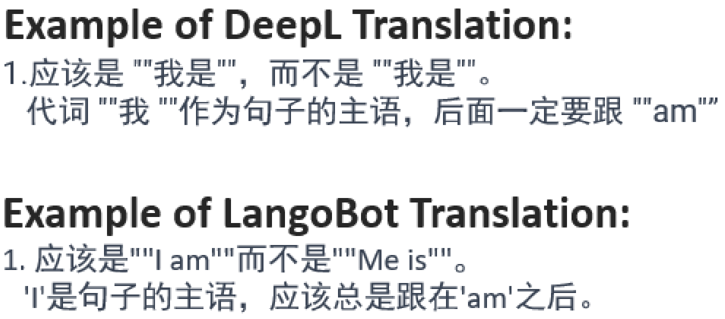
Languages	Chinese	German	French	Spanish
BLEU Score	0.380189	0.529708	0.617428	0.584771

Table 3. BLEU Scores for Various Languages

According to the BLEU scores in Table 3, three of the languages achieved scores above 0.5, demonstrating a strong similarity to DeepL translations, a recognized high standard in the industry. The comparatively lower score for Chinese translations may be attributed to its unique tokenization requirements.

Qualitative results and discussion

Below, we present an example of Chinese translations from LangoBot and DeepL. Interestingly, LangoBot's translations often retain key English phrases in their original form, which is an essential feature for English tutoring purposes. Generally, LangoBot preserved more English phrases than DeepL, leading us to conclude that it performs better for our specific educational objectives.



Example of DeepL Translation:

1.应该是 ""我是""，而不是 ""我是""。
代词 ""我 ""作为句子的主语，后面一定要跟 ""am"”

Example of LangoBot Translation:

1. 应该是""I am""而不是""Me is""。
'I'是句子的主语，应该总是跟在'am'之后。

Figure 9. Example Translations

Feedback

Quantitative results

Quantitatively, we manually assessed the quality of the generated overall feedback by evaluating its accuracy, recall, and precision in identifying users' grammar or reading comprehension weaknesses, indicated by any grammar mistakes or irrelevant responses in the conversation. These results are detailed in Table 4.

	Grammar	**Reading Comprehension**
Accuracy	0.8095	0.8095
Recall	0.9	0.625
Precision	0.75	0.8333

Table 4. Quantitative Results for the Overall Feedback

Qualitative results

We established qualitative criteria to assess the relevance and accuracy of the advice given in the feedback. The specific criteria are detailed in Table 5, with an overall normalized score of **0.8095**. A notable portion of incorrect advice stems from inaccurately identified grammar errors. For instance, the feedback '*An incorrect use of the verb 'to be*' is observed in '*Barcelona was play against Paris Saint Germain*' incorrectly identifies a grammar issue in the sentence.

Score	**Standard**
0	Mostly irrelevant or frequently provides incorrect advice (more than 4 instances)
1	Partially irrelevant or includes some incorrect advice (3-4 instances)
2	Mostly relevant or includes minor inaccuracies (1-2 instances)
3	Highly relevant and accurate, consistently offers correct advice (0 mistakes)

Table 5. Qualitative Evaluation Criteria for the Overall Feedback

Discussion

The performance of the overall feedback is somewhat lower than that of the grammar feedback, likely because GPT-4 faces greater challenges when analyzing longer texts. Nonetheless, this proficiency feedback remains a valuable tool for learners to identify their strengths and weaknesses.

6. Learnings

Due to time and cost constraints, our evaluation results may not fully represent the system's capabilities because of the insufficient data. Despite this, the results, particularly in grammar correction and translation, demonstrate GPT-4's great capability to be integrated in language learning chatbots.

For future projects, we want to train a regressor for automated scoring based on our predefined criteria, so that we can expand our evaluation dataset efficiently. Additionally, we're considering user interface improvements such as implementing hover-over translations and displaying grammar feedback in a pop-up box upon a button click, rather than clustering all features in one chat box. Moreover, as GPT shows reduced performance with longer texts, we aim to improve grammar feedback accuracy by segmenting inputs prior to processing through the GPT-4 API, potentially enhancing outcomes.

7. Individual Contributions

Xinyu

I mainly implemented the Streamlit interface and refined the prompts after Lanzhi's initial version. I was responsible for the grammar evaluation code (GPT-4 version) and 40 manual evaluations on the grammar feedback. I was also responsible for the feedback evaluation.

Lanzhi

I assisted with the interface development and wrote all prompts, continuing with the prompt engineering work. I was responsible for the translation evaluation. Initially, I managed the grammar feedback evaluation using GPT-3.5, including writing conversations and manual labelling and evaluation for 84 examples, though this was not part of the final deliverable. I also conducted 40 evaluations of grammar feedback.

References

[1] G. Tyen, M. Brenchley, A. Caines, and P. Buttery, ‘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), 2022, pp. 234–249.

[2] N. Kshetri, ‘The Economics of Generative Artificial Intelligence in the Academic Industry’, Computer, vol. 56, no. 8, pp. 77–83, 2023.