EDEN: Empathetic Dialogues for English learning

Li Siyan, Teresa Shao, Zhou Yu, Julia Hirschberg
Department of Computer Science
Columbia University
{siyan.li,ts3488,zy2461,jbh2019}@columbia.edu

Abstract

Dialogue systems have been used as conversation partners in English learning, but few have studied whether these systems improve learning outcomes. Student passion and perseverance, or *grit*, has been associated with language learning success. Recent work establishes that as students perceive their English teachers to be more supportive, their grit improves. Hypothesizing that the same pattern applies to English-teaching chatbots, we create EDEN, a robust open-domain chatbot for spoken conversation practice that provides empathetic feedback. To construct EDEN, we first train a specialized spoken utterance grammar correction model and a high-quality social chit-chat conversation model. We then conduct a preliminary user study with a variety of strategies for empathetic feedback. Our experiment suggests that using adaptive empathetic feedback leads to higher *perceived affective support*. Furthermore, elements of perceived affective support positively correlate with student grit.

1 Introduction

We study chatbots that teach languages like English (Ayedoun et al., 2020, 2015; Yang et al., 2022; Kohnke, 2023), and in particular how they can improve student persistence in learning a second language. In the language learning literature, this is referred to as *L2 grit* (Teimouri et al., 2022). High L2 grit is crucial for student well-being and success. For example, L2 grit correlates strongly with increased learning enjoyment (Elahi Shirvan et al., 2021), negatively predicts foreign language anxiety and burnout (Li and Dewaele, 2021; Wu et al., 2023), and indirectly predicts L2 achievement (Khajavy and Aghaee, 2022). While L2 grit is often framed as a personality trait, changing an individual's grit is possible (Hwang and Nam, 2021; Tang et al., 2019; Pueschel and Tucker, 2018). As a result, a promising direction is to study grit-improving interventions.

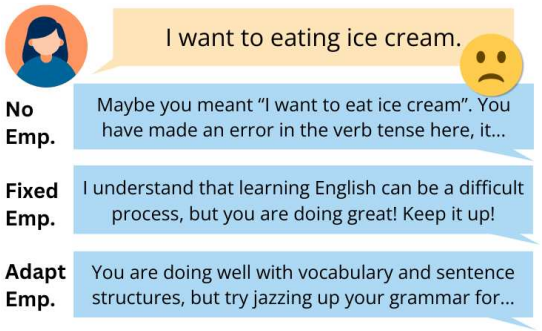

Figure 1: The three different empathetic feedback strategies in our experiments. This is a special case where the input is grammatically incorrect, so the No Empathetic Feedback condition would provide corrections.

Wu et al. (2023) examines how teachers' perceived affective support (PAS), i.e. how supportive the students perceive their teachers to be (Sakiz, 2007), influences student L2 grit in Chinese ESL medium-level learners in a college-level English class. The authors discover a strong predictive relationship between higher teacher PAS and higher student L2 grit. This further encourages teachers to exhibit warmth and empathy to improve their PAS. The study does not establish whether higher teacher PAS improves L2 grit, but we posit that it is possible.

We seek to determine whether the relationship between PAS and L2 grit extends from human teachers to English-teaching dialogue systems. That is, **does higher *chatbot* perceived affective support also increase student L2 grit?** Since perceived affective support correlates positively with empathy, an empathetic English-teaching chatbot should allow us to study this relationship. However, there is little work on incorporating empathy into open-domain English-teaching chatbots (Zhai and Wibowo, 2022).

To bridge this gap, we construct EDEN (**E**mpathetic **D**ialogues for **EN**glish learning), which is a high-quality and robust dialogue sys-

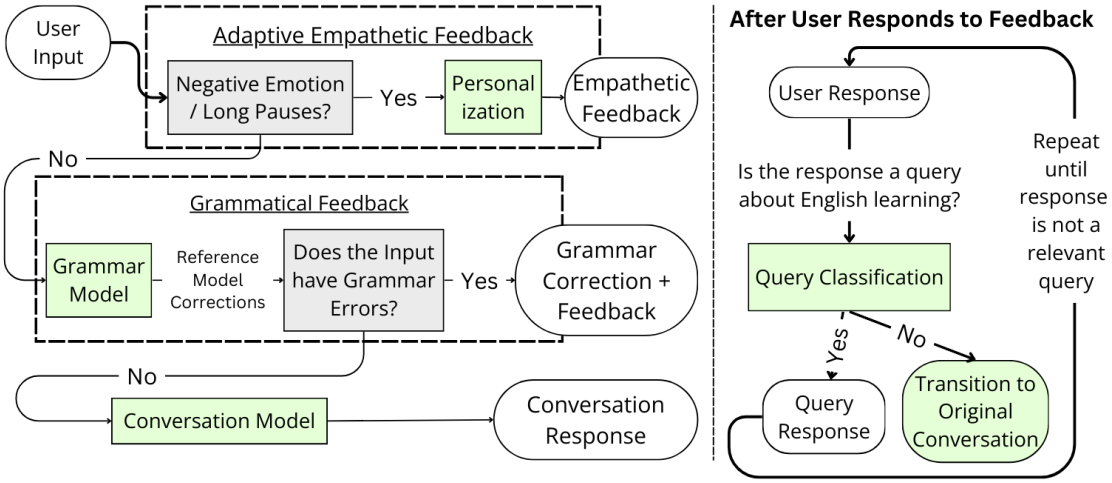

Figure 2: An overview of EDEN's architecture. We highlight several additions and improvements compared to the design by Siyan et al. (2024) in green.

tem capable of empathetic and grammatical feedback. To strengthen components of EDEN for educational spoken dialogue, we tailor our grammatical feedback for spoken utterances, build a conversation model for open-domain chitchat across multiple topics, and introduce personalization to cater to user preferences. We encourage others to expand upon and customize EDEN for research within and outside of language education.[1]

Using EDEN, we conduct a preliminary user study on how empathetic feedback mechanisms influence the chatbot's perceived affective support and user L2 grit changes. Our results suggest that the adaptive empathetic feedback strategy is the most successful in inducing high perceived affective support. This could be due to the specificity of the adaptive mechanism making users feel more thoughtfully attended to. Additionally, we discover that certain components of chatbot perceived affective support predict positive changes in L2 grit, which aligns with our hypotheses.

2 Related Work

Empathetic chatbots have been applied to counseling (DeVault et al., 2014; Trappey et al., 2022), medical assistance (Daher et al., 2020), motivation for weight management (Rahmanti et al., 2022), customer service (Xu et al., 2017), or for social and communicative needs (De Gennaro et al., 2020; Jiang et al., 2022). However, there is little work on integrating empathy into second-language education using current-day language technologies.

There have been affective English educational conversation systems. Ayedoun et al. (2020), improving upon Ayedoun et al. (2015), presents a multimodal agent for improving L2 learners' willingness to communicate. The agent carries out a pre-scripted dialogue and adopts different communicative strategies and affective backchannels to reduce learner anxiety. Shi et al. (2020) creates an empathetic spoken chatbot for pronunciation correction using an ontology. Lee et al. (2023) trains a real-world situational chatbot capable of providing feedback. Park et al. (2022) incorporates persona-based conversation capabilities into a humanoid robot to make language practice easier for anxiety-prone individuals. Nonetheless, none of these chatbots account for student emotions. The learner-emotion-aware systems, on the other hand, are often not conversational (Lin et al., 2015; Wu et al., 2022; Santos et al., 2016).

Empathetic strategies have been employed in other forms of learning. Affective AutoTutor (D'mello and Graesser, 2013) responds with emotional statements to regulate negative student emotions in physics tutoring. Litman and Forbes-Riley (2014) modifies a spoken physics education system (Litman and Silliman, 2004) that adapts to user affective states to identify and respond to real-time user disengagement. Other affective tutoring systems often offer hints to resolve student frustration (Hasan et al., 2020; Fwa, 2018; Lin et al., 2014). We postulate that these approaches may be proven

[1]We release all data, code, and model checkpoints under an open-source license here: https://github.com/siyan-sylvia-li/EDEN

effective for language learning as well.

Prior work has explored the relationship between chatbot usage and L2 learner experience. Han (2021) reports Korean EFL learners experiencing reduced language anxiety and enhanced English-learning interests when using English chatbots. AI-mediated discussions have also been more effective than face-to-face discussions for increasing L2 learners' willingness to communicate (Fathi et al., 2024). However, there is no systematic study on L2 grit changes in the chatbot context.

3 Chatbot Design

While using dialogue systems for English education has become popular, there are only a few fully open-source systems using state-of-the-art methods. In our work, we construct an extensible and robust spoken dialogue system as a conversation partner. Figure 2 shows an overview of EDEN, which makes several key improvements over an empathetic chatbot architecture proposed by Siyan et al. (2024).[2] For each turn, the user input is first analyzed for negative sentiment and prolonged pauses. If these signals are captured, corresponding empathetic feedback is synthesized. Otherwise, a grammatical feedback module utilizes a grammar correction model to locate grammar errors and construct feedback using templates. We further devise a grammar correction hierarchy to not overwhelm users (Appendix E). When the user utterance does not trigger grammatical feedback, the conversation proceeds as normal through a conversation language model.

After the user receives feedback, they occasionally have follow-up queries. We resort to ChatGPT for resolving these queries if they are relevant to English learning. We use ChatGPT for this purpose rather than our conversation model because responding to user queries about the feedback would be out of the scope of our conversation model. Instead, we employ a transition module (Appendix H) to continue the original conversation after ChatGPT responses. Additional design choices are informed by our chatbot design survey conducted on Twitter / X with more than 450 responses (Appendix B).

We make several innovations in our system design to strengthen EDEN for *spoken*, *educational*, and *open-domain* English-practice dialogue. We discuss how we create a grammar feedback model for spoken utterances (Section 3.1) and an open-domain chit-chat model (Section 3.2), detail the construction of adaptive empathetic feedback (Section 3.3), and specify how users can customize EDEN according to their needs (Section 3.4).

3.1 Tailored Grammatical Feedback for Spoken Utterances

Prior work on grammar corrections focuses on written texts rather than dialogue (Yasunaga et al., 2021; Bryant and Briscoe, 2018; Katinskaia and Yangarber, 2023). Yuan et al. (2022) presents a model trained on dialogue utterances, but the collected dataset focuses on written dialogue, which can differ from spoken dialogue. There is a lack of dialogue grammar correction datasets for spoken conversations. We bridge this gap by training a grammar correction model specialized for transcribed spoken utterances. Since the grammar correction process occurs for every user utterance, and constant GPT-4 API queries are costly and inefficient, we use a local model instead of GPT-4.

3.1.1 Grammar Correction Data

To train such models, we utilize a set of Whisper-Medium-transcribed audio clips collected from an existing English learning chatbot platform (Li et al., 2022) proposed by Siyan et al. (2024). The speakers of these clips are native Mandarin speakers, which is representative of our target users. We adopt this set for training a grammar correction model specialized for spoken utterances.

Since there is a lack of ground-truth grammar corrections for these utterances, and language models such as GPT-4 are effective in grammar error correction tasks (Kobayashi et al., 2024; Coyne et al., 2023), we use GPT-4-Turbo to generate grammar corrections for these transcriptions. In total, we produce 7,894 sentence-correction pairs.

3.1.2 Model Training and Evaluation

We train two models, a Llama-2 7B model and a Flan-T5-XL model, on the same data. Both models are trained on a single GPU using Parameter-Efficient Fine-Tuning (Mangrulkar et al., 2022). See further training details in Appendix C.

Due to a lack of ground-truth data, we again use GPT-4-Turbo to compare model performance on the validation set. For sentences where the two models disagree, we prompt GPT-4-Turbo to assess whether the model corrections are grammatically

[2]We further discuss the original chatbot design in Appendix A and more information can be found at the original paper's site https://github.com/siyan-sylvia-li/adaptive_empathetic_BEA2024

Original	Flan-T5-XL	Llama-2
It's not famous, but I really love it because it's my hometown, and I live here until now with my family and friends.	It's not famous, but I really love it because it's my hometown, and I've lived here until now with my family and friends.	It's not famous, but I really love it because it's my hometown and I still live here with my family and friends.
Dry weather will give a dry skin and some illness comes from it.	Dry weather causes dry skin, and some illnesses result from it.	Dry weather will give dry skin and lead to some illnesses.

Table 1: Example grammar corrections generated by our two models.

	#L-Correct	#L-Wrong
#F-Correct	61.6%	11.1%
#F-Wrong	12.9%	14.4%

Table 2: Percentages of validation set transcribed utterances that **Llama-2** generates valid and invalid corrections for (#**L**-Correct and #**L**-Wrong), as well as the number of utterances that **Flan-T5-XL** provides valid and invalid corrections for (#**F**-Correct and #**F**-Wrong), as measured by **GPT-4**.

Study	Win%	Lose%	Tie%
Study #1, *All*	**46.3**	35.8	17.9
Study #1, *Maj*	**45.0**	34.5	20.5
Study #2, *All*	39.3	**40.3**	20.4
Study #2, *Maj*	40.1	**41.4**	18.5

Table 3: **Llama-2 vs. Flan-T5** win, lose, and tie rates from the two **human subject** studies. *All* indicates including all sentences, and *Maj* indicates the results when only including sentences with a preferred majority (more than 50% of participants voted for one option) among participants.

correct. We report the judgment results on these utterances in Table 2. We find that the two models usually both provide valid corrections, with the Llama-2 model slightly out-performing.

For a more robust evaluation, we conduct two human-subject studies by recruiting participants from Prolific. Our goal is to compare the two models *when they differ*, under two conditions: (i) when at least one reports a grammatical error, which essentially assesses their *precision* (Study #1), and (ii) when an expert identifies a grammatical mistake in the original utterance, which essentially assesses their *recall* (Study #2). For Study #1, a random subset of 40 medium-length sentences from the validation set. For Study #2, we curate 31 such sentences that are grammatically incorrect. In both studies, six native English speaker participants, paid at $15 per hour, compare the corrections for each sentence from the two models. The participants are asked to select the better correction, defined as (1) minimally changing the original sentence, (2) retaining the original meaning, and (3) grammatically correct. Each participant evaluated 20 sentences for Study #1 and 31 sentences for Study #2.

Since both models tend to generate valid corrections, participant preferences vary, resulting in low inter-annotator agreement. For the first study, the Fleiss' kappas for two 20-sentence batches are 0.310 and 0.301 respectively (fair agreement). For the second study, the kappa is 0.139 (limited agreement). Therefore, we additionally evaluate participant preferences for sentences with a majority of participants agreeing (Table 3).

While participants prefer Llama-2 for randomly selected transcripts, they slightly prefer Flan-T5-XL for erroneous sentences. This could be due to Llama-2 providing higher-quality rewrites to grammatically correct sentences. Considering that Llama-2 is generally preferred in Study #1 and the differences in Study #2 are minor, we choose the fine-tuned Llama-2 as the grammar model.

3.2 Open-Domain Conversation Model

Previous English chatbots designed for speaking practice focus on delivering course content in formal English training (Du and Daniel, 2024). Here, we are targeting users who are learning English out of interest. Therefore, EDEN must engage users in interesting conversations to improve their experience and reduce their language anxiety (Von Worde, 2003). We create a conversation model capable of discussing various topics to accommodate user interest. Although ChatGPT or GPT-4o is a convenient choice, GPT-4o can be too slow for chit-chat and is not widely accessible in China, where some of our recruited users are located.

We adapt a data synthesis pipeline with persona-based prompting (Li et al., 2023) to support social chit-chat while retaining its strength in generating naturalistic and accessible conversation responses and taking initiatives. Specifically, we use everyday topics (e.g. favorite cuisine), remove some extraneous constraints in the original pipeline, and adjust prompts after analyzing preliminary outputs.

The following broad topic areas are used: *food, books, movies, TV shows, music, hobbies, and English learning*. We further identify 243 relevant topics within these broad areas (Appendix F.1).

We ask ChatGPT to generate 10 two-party conversations per topic. One of the personas is generic (often assumed to be American by ChatGPT), and the other is someone whose first language is not English. We originally used a hypothetical Chinese college student as the second persona to be consistent with prior work, but the generation diversity was problematically low (Appendices F.2 and F.3).

Upon further examination of the generated data, we discover some low-quality generated conversations, which we address through filtering. These lower-quality conversations often feature one conversation party making assumptions about the other party. See details about these quality issues in generated conversations, as well as the data filtering process, in Appendix F.4. After filtering, 1,227 conversations remain. The conversations have an average length of 8.35 turns. We then fine-tune a Llama-2 model on this conversation data (See Appendix D for details). Please see the distribution of topics among the conversations in Appendix F.5.

3.3 Adaptive Empathetic Feedback

Following Siyan et al. (2024), EDEN's empathetic feedback mechanism triggers when the system registers signals of user distress such as heightened negative affect or prolonged pauses. A ChatGPT prompt, optimized through the DSPy framework (Khattab et al., 2023), is used to produce a piece of feedback from past user utterances. Generally, the feedback (i) sounds empathetic and colloquial, (ii) includes examples and actionable feedback. Since the generated feedback can still sound overly formal, we use additional rewrite prompts to shorten the feedback and reduce its formality.

3.4 Personalization Feature

We notice in our design survey that users have a variety of preferences. Two design aspects that reflect such diversity are whether to include Mandarin translations of chatbot utterances and the length of chatbot feedback. We thus allow users to customize EDEN by including personalization questions in our experiment flow before any conversations:
Q1: Do you want Mandarin translations of the chatbot utterances? (Yes / No)

If the participant selects Yes, each chatbot utterance is translated into Mandarin using ChatGPT.
Q2: How would you like the chatbot feedback? (Succinct / Details & examples / No preference)

Adaptive empathetic feedback utterances are customized through prompting using user responses. The original generated feedback is used if the participant has no preference (Appendix G).

4 User Study: Empathetic Feedback

We recruited 31 native Mandarin speakers from the internet (15) and the authors' home institution (16). The internet participants were not compensated, while the participants from the author's institution received $15 Amazon gift cards. Our IRB-approved study intends to answer these research questions:
RQ1: Does adaptive, empathetic feedback in an English-teaching chatbot result in higher perceived affective support (PAS)?
RQ2: Does higher chatbot PAS correlate to positive changes in L2 grit?

4.1 Experimental Conditions

Our participants are assigned to one of three experimental conditions sequentially:

1. *No Empathetic Feedback (**None**).*
2. ***Fixed** Empathetic Feedback:* The empathetic feedback is randomly selected from a predefined list of generic empathetic phrases. See the fixed empathetic responses in Appendix I.
3. ***Adaptive** Empathetic Feedback:* The empathetic feedback is generated through ChatGPT using prompts from Siyan et al. (2024). We personalize this feedback.

All other components of EDEN (i.e. grammatical feedback, conversation, etc.) are held constant across conditions. By defining the conditions as such, we can more rigorously test the effect of the presence and the different types of empathetic feedback on student L2 grit and chatbot PAS. We hypothesize that:

H1: Both Fixed and Adaptive conditions improve chatbot PAS, but Adaptive is more effective.
H2: Higher chatbot PAS would correlate to positive changes in L2 grit.

4.2 Experimental Procedure

The participants first complete a pre-survey about their English proficiency and L2 grit (Teimouri et al., 2022). They then proceed to converse with EDEN for at least three conversations after completing the short personalization questionnaire. Upon completion of the chatbot interaction phase, the participants evaluate their experience, the chatbot's PAS, and their L2 grit in a post-survey. We use the same adapted chatbot PAS survey (Siyan et al., 2024) and the L2 grit survey (Teimouri et al., 2022). All questions are five-item Likert-scale questions presented in both English and Mandarin.

Overloading the question codes to be the reported values for their questions, we further define:

$$PAS = \frac{ENC + LIST + CARE + APP}{4}$$
$$\Delta L2.k = L2.k_{post} - L2.k_{pre}$$

where $L2.k_{post}$ and $L2.k_{pre}$ refer to the reported values for the L2-grit-related question $L2.k$ in the post-survey and the pre-survey, respectively. Since the survey has items 2, 4, 7, and 8 reverse-coded, we compute the total change in L2 grit as such:

$$\begin{aligned}\Delta L2_{Total} = \Delta L2.1 - \Delta L2.2 + \Delta L2.3 \\ - \Delta L2.4 + \Delta L2.5 + \Delta L2.6 \\ - \Delta L2.7 - \Delta L2.8 + \Delta L2.9\end{aligned}$$

5 Results and Discussion

Our participants display intermediate self-reported English proficiency. They have studied English for an average of 15.9 years. Since Chinese citizens tend to start learning English at a young age, this number is not out of the ordinary. Their average scores for IELTS and TOEFL are 6.7 and 110.6, respectively. Most participants speak more Mandarin than English in their everyday lives.

On average, the participants conversed with our chatbot for 31.19 turns[3]. They received 1.57 grammatical feedback during the interactions, and participants under Fixed and Adaptive conditions received 4.42 and 2.67 empathetic feedback, respectively. The top three selected topics are food, English learning, and books.

[3]We were not able to locate two None condition participants' conversation data due to experiment ID mismatch.

Question Code: Question Text
QUAL: How was the conversation quality? **CONF:** Do you feel more confident after conversing with the chatbot? **USE:** Do you think the chatbot's grammar feedback is useful?
ENC: The chatbot encourages me when I am having difficulties in the conversation. **LIST:** The chatbot listens to me when I have something to say. **CARE:** My opinion matters to the chatbot. **APP:** The chatbot recognizes and appreciates when I am good at something.
L2.1: I am a diligent English language learner. **L2.2:** My interests in learning English change from year to year. **L2.3:** When it comes to English, I am a hard-working learner. **L2.4:** I think I have lost my interest in learning English. **L2.5:** Now that I have decided to learn English, nothing can prevent me from reaching this goal. **L2.6:** I will not allow anything to stop me from my progress in learning English. **L2.7:** I am not as interested in learning English as I used to be. **L2.8:** I was obsessed with learning English in the past but have lost interest recently. **L2.9:** I put much time and effort into improving my English language weaknesses.

Table 4: Select questions used for measuring general conversation quality (top), PAS (middle), and L2 grit (bottom) in the pre- and post-surveys. We provide additional survey details in Appendix J.2.

Due to our participants having intermediate English proficiency on average, some participants do not trigger the empathetic feedback module. We therefore reassign participants who did not trigger empathetic feedback to None condition. After this reassignment, we have 17 participants for the None condition, six for Fixed, and eight for Adaptive.

Overall, the participants consider the conversations to be moderate-to-high quality ($\overline{\text{QUAL}} = 3.39$). They experience some confidence boost post-interaction ($\overline{\text{CONF}} = 3.39$), and they find the grammar feedback useful ($\overline{\text{USE}} = 3.52$). Some conversation quality ratings were negatively affected by network errors during experiments.

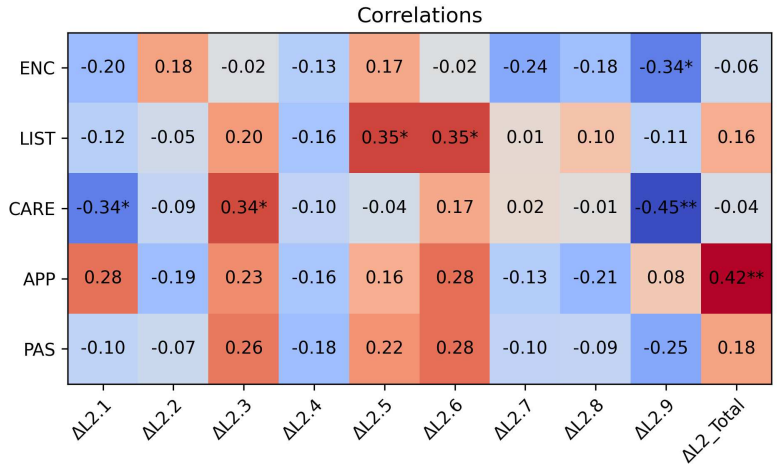

Figure 3: Correlations between different measures for chatbot PAS and L2 grit changes. **= $p < 0.05$, *= $p < 0.1$.

Δ for	**L2.1**	**L2.2**	**L2.3**	**L2.4**	**L2.5**	**L2.6**	**L2.7**	**L2.8**	**L2.9**	**L2**$_{Total}$
None	0.05	-0.24	-0.05	0.29	-0.12	-0.24	0.24	0.47	0.47	-0.64
Fixed	**0.83**	0.67	**0.33**	-0.17	**0.17**	0.00	**-0.17**	-0.33	**0.83**	**2.17**
Adaptive	0.25	**-0.50**	0.13	**-0.25**	0.13	**0.25**	-0.13	**-0.38**	0.13	2.13

Table 5: Average L2 grit changes for the different experimental conditions. Note that items 2, 4, 7, and 8 are reverse-coded, so lower would be better for these items.

	ENC	**LIST**	**CARE**	**APP**	**PAS**
None	3.53	**4.12**	**4.00**	3.47	3.78
Fixed	3.83	2.83	3.00	3.67	3.33
Adap.	**4.38**	4.00	3.88	**4.38**	**4.16**

Table 6: Average PAS questionnaire results for the different experimental conditions.

5.1 Causal Relationship between Empathetic Feedback and PAS

We present the post-survey results for chatbot PAS for the different conditions in Table 6. Adaptive outperforms Fixed for all PAS-related metrics. This is expected, as a generic phrase is unlikely to elicit as much perceived empathy as a tailored, adaptive piece of feedback. Furthermore, using fixed phrases may be perceived as more unnatural in a conversation than using a personalized response, making the participants feel not listened to. The Adaptive condition results in the highest PAS in the pilot study, although this dominance does not persist across different items. Users feel more listened to and that their opinions matter more under the None condition potentially for a similar reason; the current potentially unnatural transition between dialogue and feedback content can lead to the users perceiving a lack of chatbot attention. Adaptive performs the best in encouragement and appreciation, and both Adaptive and Fixed are better than None here. This could indicate that EDEN's empathetic feedback mechanism correctly identifies and addresses participant struggles. Including praises in the empathetic feedback pipeline likely contributes to a higher appreciation rating. Our adaptive condition also causes a higher PAS rating compared to what Siyan et al. (2024) reported, which is 3.27, highlighting our improvement.

These results validate the first hypothesis. We postulate that by making EDEN's transition between conversation and feedback more seamless, we can enhance perceived affective support further by helping users feel better attended to.

5.2 Correlation between PAS and L2 Grit

Table 5 records the average L2 grit changes per condition. We notice that the None condition never achieves the most positive L2 grit changes. Meanwhile, although the Fixed condition is associated with the lowest PAS, it achieves the highest overall L2 grit changes, slightly above Adaptive.

Pearson's correlation is used to evaluate the relationship between various components of PAS and changes in L2 grit (Figure 3). Our results show-

case some components of PAS being weak to intermediate predictors for positive L2 grit changes. Specifically, perceived chatbot appreciation correlates positively with changes in total L2 grit, and users feeling their opinions matter predicts positive changes in self-perception of being hard-working. Additionally, users feeling that they are listened to is correlated with increased self-determination.

We identify counter-intuitive results that can be attributed to our small sample size. By the L2 grit questionnaire definition, ΔL2.1 and ΔL2.3 should be positively correlated. However, users feeling their opinions matter positively correlates with one and not the other. Similarly, it negatively correlates with ΔL2.9, the self-perception of putting much effort into improving English skills. Higher PAS still weakly correlates with positive L2 grit changes, suggesting that our results partially align with Wu et al. (2023) and supporting our second hypothesis.

5.3 Additional Correlations

PAS and Conversation Quality: During the user study, we noticed that negative bot interactions can reduce PAS. We are therefore curious about how PAS relates to conversation quality ratings. We find significant positive correlations between PAS and the conversation quality measures (Appendix L.1). This result suggests using PAS as a reliable conversation quality measure for social chatbots.

English Proficiency and L2 Grit Changes: We discover that English proficiency does not significantly correlate with a total of L2 grit changes. This indicates that, in our pilot study, being more proficient does not preclude users from having higher L2 grit after chatbot interactions. If this result generalizes, English learners from all levels could benefit from chatbot interventions that improve grit.

PAS and L2 Grit: Directly reproducing Wu et al. (2023) results, we examine the correlation between various PAS measures and *L2 grit* in the post-survey (Appendix L.2). We find that higher perceived appreciation is an intermediate-strength predictor for higher L2 grit in the post-survey. This result partially validates the generalizability of Wu et al. (2023) results to chatbot settings, since only perceived appreciation serves as a sufficient predictor, and overall PAS has no significant correlations with any of the L2 grit measures.

5.4 User Feedback

Users generally appreciate the quality of recommendations made by EDEN and the naturalness of chatbot responses. One user commented that EDEN provides recommendations highly tailored to their preferences. Several users commended how engaging the conversations were. Another user acknowledged the benefits of using a chatbot as a language practice partner: *"Notably, I felt more at ease communicating with the AI than with a human, as there is often a fear of judgment regarding one's speaking abilities."* A few users indicated excitement about trying EDEN in the future as a commercial product. However, some participants dislike the grammar feedback since they already have high English proficiency and do not require the more basic feedback.

Participants assigned to both empathetic conditions enjoyed the chatbot's supportiveness. One of the participants under the adaptive condition said, "*I was rather surprised when I received the encouraging feedback but in a good way.*"

Currently, EDEN has little capability beyond social chitchat and providing grammatical and empathetic responses. Therefore, it would fail when users request their English skills to be evaluated (e.g. "How good do you think my English speaking skill is?" or "Could you evaluate my English skill") or query the number of turns in the current conversation. These requests could signal participants placing trust in EDEN's capabilities. Another failure mode occurs when the chatbot uses vocabulary beyond the users' comprehension. Future work can address this by developing additional functionalities and user-adaptive mechanisms for vocabulary choice. We provide example conversations in Appendix K.

6 Conclusion

In this work, we build EDEN, a robust open-domain empathetic English-teaching chatbot tailored for spoken conversations. We then use it to verify whether results from Wu et al. (2023) extend to perceived affective support of chatbots in addition to teachers. Our initial user study reveals that higher perceived affective support of our chatbot correlates positively with changes in student L2 grit. We additionally showcase that adaptive, empathetic feedback surpasses fixed and no empathetic feedback in enhancing chatbot perceived affective support. Our work serves as a first step in exploring dialogue system interventions for boosting L2 grit.

7 Limitations

Our human evaluation has several limitations. Due to a lack of convenience samples, we had to recruit from the internet and the authors' home institution which is a university in the United States. This poses a sampling bias since individuals recruited through these channels tend to have at least moderate English proficiency; thus the diversity in our sample is inherently limited. Furthermore, since we were not compensating our online user study participants, it was difficult to retain these participants. Another issue with our experimental results lies in the imbalance of participants in the different experimental conditions. Since our empathetic intervention is based on participants exhibiting distress or prolonged pauses in speech, it is inherently more frequently triggered with more beginner-level participants, which does not align with the typical Mandarin native-speaker population residing in the US. We have attempted to expand our experiment to Mainland Chinese users, but they tend to experience connection issues to our servers. We currently only target native Mandarin speakers because the system from Siyan et al. (2024) is developed from Mandarin speaker data and focuses on native Mandarin speakers. We may experiment with expanding to all English-as-a-second-language individuals in the future.

In terms of implementation, areas such as Text-to-Speech synthesis could use further improvement, as the perception of empathy often hinges upon the tone of the voice. Furthermore, as discussed in the user feedback section, some users have asked for assessments of English skills and the definition of the assessments, as well as how long the conversation has been, which we do not have a standardized set of guidelines for. Future work could incorporate further modularization with different functionalities. There were also intermittent server connection issues on the users' ends which negatively impacted their experience.

8 Ethical Considerations

As we observe in the study, some users would assign authority to the chatbot and request feedback. When the feedback is not truthful or not sufficiently encouraging, the users may experience negative emotions as a result. We must additionally ensure minimizing and eliminating any harmful content the chatbot could produce when interacting with users. Users anthropomorphizing our chatbot is beneficial for perceiving empathy from the chatbot, but we should strike a delicate balance to avoid users becoming emotionally dependent or attached.

Using automated systems for education purposes might raise concerns about these pedagogical systems replacing teachers. We would like to state that our chatbot is intended to be a conversation practice partner outside of the classroom, and is not a replacement for human instruction.

References

Emmanuel Ayedoun, Yuki Hayashi, and Kazuhisa Seta. 2015. A conversational agent to encourage willingness to communicate in the context of english as a foreign language. *Procedia Computer Science*, 60:1433–1442.

Emmanuel Ayedoun, Yuki Hayashi, and Kazuhisa Seta. 2020. Toward personalized scaffolding and fading of motivational support in l2 learner–dialogue agent interactions: an exploratory study. *IEEE Transactions on Learning Technologies*, 13(3):604–616.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Christopher Bryant and Ted Briscoe. 2018. Language model based grammatical error correction without annotated training data. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 247–253, New Orleans, Louisiana. Association for Computational Linguistics.

Leshem Choshen, Matanel Oren, Dmitry Nikolaev, and Omri Abend. 2021. Serrant: a syntactic classifier for english grammatical error types. *arXiv preprint arXiv:2104.02310*.

Steven Coyne, Keisuke Sakaguchi, Diana Galvan-Sosa, Michael Zock, and Kentaro Inui. 2023. Analyzing the performance of gpt-3.5 and gpt-4 in grammatical error correction. *arXiv preprint arXiv:2303.14342*.

Karl Daher, Jacky Casas, Omar Abou Khaled, and Elena Mugellini. 2020. Empathic chatbot response for medical assistance. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*, pages 1–3.

Mauro De Gennaro, Eva G Krumhuber, and Gale Lucas. 2020. Effectiveness of an empathic chatbot in combating adverse effects of social exclusion on mood. *Frontiers in psychology*, 10:495952.

David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, et al. 2014. Simsensei kiosk: A virtual human interviewer

for healthcare decision support. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pages 1061–1068.

Sidney D'mello and Art Graesser. 2013. Autotutor and affective autotutor: Learning by talking with cognitively and emotionally intelligent computers that talk back. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(4):1–39.

Jinming Du and Ben Kei Daniel. 2024. Transforming language education: A systematic review of ai-powered chatbots for english as a foreign language speaking practice. *Computers and Education: Artificial Intelligence*, 6:100230.

Majid Elahi Shirvan, Tahereh Taherian, Mojdeh Shahnama, and Elham Yazdanmehr. 2021. A longitudinal study of foreign language enjoyment and l2 grit: A latent growth curve modeling. *Frontiers in Psychology*, 12:720326.

Jalil Fathi, Masoud Rahimi, and Ali Derakhshan. 2024. Improving efl learners' speaking skills and willingness to communicate via artificial intelligence-mediated interactions. *System*, 121:103254.

Hua Leong Fwa. 2018. An architectural design and evaluation of an affective tutoring system for novice programmers. *International Journal of Educational Technology in Higher Education*, 15(1):1–19.

Daeun Han. 2021. An analysis of korean efl learners' experience on english classes using ai chatbot. *Robotics & AI Ethics*, 6(3):1–9.

Muhammad Asif Hasan, Nurul Fazmidar Mohd Noor, Siti Soraya Binti Abdul Rahman, and Mohammad Mustaneer Rahman. 2020. The transition from intelligent to affective tutoring system: a review and open issues. *IEEE Access*, 8:204612–204638.

Mae-Hyang Hwang and JeeEun Karin Nam. 2021. Enhancing grit: Possibility and intervention strategies. *Multidisciplinary perspectives on grit: Contemporary theories, assessments, applications and critiques*, pages 77–93.

Qiaolei Jiang, Yadi Zhang, and Wenjing Pian. 2022. Chatbot as an emergency exist: Mediated empathy for resilience via human-ai interaction during the covid-19 pandemic. *Information processing & management*, 59(6):103074.

Anisia Katinskaia and Roman Yangarber. 2023. Grammatical error correction for sentence-level assessment in language learning. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 488–502, Toronto, Canada. Association for Computational Linguistics.

Gholam Hassan Khajavy and Elham Aghaee. 2022. The contribution of grit, emotions and personal bests to foreign language learning. *Journal of Multilingual and Multicultural Development*, pages 1–15.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2023. Dspy: Compiling declarative language model calls into self-improving pipelines. *arXiv preprint arXiv:2310.03714*.

Masamune Kobayashi, Masato Mita, and Mamoru Komachi. 2024. Large language models are state-of-the-art evaluator for grammatical error correction. *arXiv preprint arXiv:2403.17540*.

Lucas Kohnke. 2023. L2 learners' perceptions of a chatbot as a potential independent language learning tool. *International Journal of Mobile Learning and Organisation*, 17(1-2):214–226.

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

Chengchen Li and Jean-Marc Dewaele. 2021. How classroom environment and general grit predict foreign language classroom anxiety of chinese efl students. *Journal for the Psychology of Language Learning*, 3(2):86–98.

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*, pages 451–455.

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. *arXiv preprint arXiv:2309.16804*.

Kai-Hui Liang, Sam Davidson, Xun Yuan, Shehan Panditharatne, Chun-Yen Chen, Ryan Shea, Derek Pham, Yinghua Tan, Erik Voss, and Luke Fryer. 2023. ChatBack: Investigating methods of providing grammatical error feedback in a GUI-based language learning chatbot. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 83–99, Toronto, Canada. Association for Computational Linguistics.

Hao-Chiang Koong Lin, Ching-Ju Chao, and Tsu-Ching Huang. 2015. From a perspective on foreign language learning anxiety to develop an affective tutoring system. *Educational Technology Research and Development*, 63:727–747.

Hao-Chiang Koong Lin, Chih-Hung Wu, and Ya-Ping Hsueh. 2014. The influence of using affective tutoring system in accounting remedial instruction on learning performance and usability. *Computers in Human Behavior*, 41:514–522.

Diane Litman and Kate Forbes-Riley. 2014. Evaluating a spoken dialogue system that detects and adapts to user affective states. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 181–185.

Diane Litman and Scott Silliman. 2004. Itspoke: An intelligent tutoring spoken dialogue system. In *Demonstration papers at HLT-NAACL 2004*, pages 5–8.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft.

Chanjun Park, Yoonna Jang, Seolhwa Lee, Sungjin Park, and Heuiseok Lim. 2022. FreeTalky: Don't be afraid! conversations made easier by a humanoid robot using persona-based dialogue. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1242–1248, Marseille, France. European Language Resources Association.

Andrew Pueschel and Mary L Tucker. 2018. Achieving grit through the growth mindset. *Journal of Instructional Pedagogies*, 20.

Annisa Ristya Rahmanti, Hsuan-Chia Yang, Bagas Suryo Bintoro, Aldilas Achmad Nursetyo, Muhammad Solihuddin Muhtar, Shabbir Syed-Abdul, and Yu-Chuan Jack Li. 2022. Slimme, a chatbot with artificial empathy for personal weight management: system design and finding. *Frontiers in Nutrition*, 9:870775.

Gonul Sakiz. 2007. *Does teacher affective support matter? An investigation of the relationship among perceived teacher affective support, sense of belonging, academic emotions, academic self-efficacy beliefs, and academic effort in middle school mathematics classrooms*. Ph.D. thesis, The Ohio State University.

Olga C Santos, Mar Saneiro, Jesus G Boticario, and María Cristina Rodriguez-Sanchez. 2016. Toward interactive context-aware affective educational recommendations in computer-assisted language learning. *New Review of Hypermedia and Multimedia*, 22(1-2):27–57.

Nuobei Shi, Qin Zeng, and Raymond Lee. 2020. The design and implementation of language learning chatbot with xai using ontology and transfer learning. *arXiv preprint arXiv:2009.13984*.

Silero. 2021. Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier. https://github.com/snakers4/silero-vad.

Li Siyan, Teresa Shao, Zhou Yu, and Julia Hirschberg. 2024. Using adaptive empathetic responses for teaching english. *arXiv preprint arXiv:2404.13764*.

Xin Tang, Ming-Te Wang, Jiesi Guo, and Katariina Salmela-Aro. 2019. Building grit: The longitudinal pathways between mindset, commitment, grit, and academic outcomes. *Journal of youth and adolescence*, 48:850–863.

Yasser Teimouri, Luke Plonsky, and Farhad Tabandeh. 2022. L2 grit: Passion and perseverance for second-language learning. *Language Teaching Research*, 26(5):893–918.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Amy JC Trappey, Aislyn PC Lin, Kevin YK Hsu, Charles V Trappey, and Kevin LK Tu. 2022. Development of an empathy-centric counseling chatbot system capable of sentimental dialogue analysis. *Processes*, 10(5):930.

Renee Von Worde. 2003. Students' perspectives on foreign language anxiety. *Inquiry*, 8(1):n1.

Chih Hung Wu, Hao-Chiang Koong Lin, Tao-Hua Wang, Tzu-Hsuan Huang, and Yueh-Min Huang. 2022. Affective mobile language tutoring system for supporting language learning. *Frontiers in Psychology*, 13:833327.

Wangjiao Wu, Yabing Wang, and Ruifang Huang. 2023. Teachers matter: exploring the impact of perceived teacher affective support and teacher enjoyment on l2 learner grit and burnout. *System*, 117:103096.

Anbang Xu, Zhe Liu, Yufan Guo, Vibha Sinha, and Rama Akkiraju. 2017. A new chatbot for customer service on social media. In *Proceedings of the 2017 CHI conference on human factors in computing systems*, pages 3506–3510.

Hyejin Yang, Heyoung Kim, Jang Ho Lee, and Dongkwang Shin. 2022. Implementation of an ai chatbot as an english conversation partner in efl speaking classes. *ReCALL*, 34(3):327–343.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2021. Lm-critic: Language models for unsupervised grammatical error correction. *arXiv preprint arXiv:2109.06822*.

Xun Yuan, Derek Pham, Sam Davidson, and Zhou Yu. 2022. ErAConD: Error annotated conversational dialog dataset for grammatical error correction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 76–84, Seattle, United States. Association for Computational Linguistics.

Chunpeng Zhai and Santoso Wibowo. 2022. A systematic review on cross-culture, humor and empathy dimensions in conversational chatbots: The case of second language acquisition. *Heliyon*.

A Chatbot Design from Siyan et al. (2024): More Details

Siyan et al. (2024) proposes a novel adaptive and empathetic English-teaching chatbot. The chatbot detects heightened negative emotions and prolonged pauses in student speech using an existing wav2vec 2.0 speech emotion detection model (Baevski et al., 2020) and a voice activity detection toolkit (Silero, 2021). When these signals are captured, ChatGPT is used to generate empathetic feedback using the past three student utterances. The authors use DSPy (Khattab et al., 2023) to optimize their ChatGPT prompt such that the resulting colloquial feedback sounds empathetic and contains specific examples.

In addition to providing empathetic feedback, the chatbot offers grammatical feedback on student utterances using a fine-tuned Llama-2 (Touvron et al., 2023) model for grammar correction. The grammar correction model is trained on the ErAConD dataset (Yuan et al., 2022), which contains written dialogue utterances and their expert grammar corrections. The SERRANT (Choshen et al., 2021) package is then used to locate the specific grammatical errors by comparing the generated correction and the original sentence. The grammatical feedback is a combination of a rephrase (e.g. *Maybe you meant "had" rather than "has"*) and a template-based explanation for each error type (Liang et al., 2023).

A transition module is designed to aid smoother transitions between different chatbot system components. It connects empathetic or grammatical feedback to the original conversation. The module classifies whether student utterances constitute an English-learning-related query using pre-written rules. If a student utterance is classified as a relevant query, the system prompts ChatGPT to generate an answer for the query, otherwise, a randomly selected pre-defined connector phrase is used to transition back to the original conversation directly.

B Chatbot Design Survey

We polled 456 Mandarin-speaking users on Twitter / X for their opinions on designing an empathetic English-teaching chatbot. The survey contains items regarding features not included in Siyan et al. (2024), such as Mandarin translations for chatbot utterances, as well as items similar to the original survey.

We present the questions and the responses from the survey below.

1. How do you like the tone of your English teacher's feedback to be?
   (a) **Colloquial (80%)**
   (b) Formal (20%)
2. How long should teacher feedback be?
   (a) 1 - 2 sentences (21.3%)
   (b) **2 - 3 sentences (52.6%)**
   (c) 3 - 4 sentences (13.2%)
   (d) 4+ sentences (12.9%)
3. If you made a mistake, how would you like your errors to be corrected? Select all that apply.
   (a) Correct your errors directly (38.8%)
   (b) Help you self-correct your errors using questions (43.6%)
   (c) **Give you examples such that you can learn from these examples and avoid making the same errors again (80.5%)**
4. What does an ideal encouraging feedback from English teachers look like? Select all that apply.
   (a) Give you encouragement, such as "You are doing great!" or "I am proud of you!" (38.2%)
   (b) Tell you what you are good at in spoken English (41.9%)
   (c) Tell you what you can do to improve your spoken English (72.4%)
   (d) **Tell you how to improve your spoken English through examples (78.7%)**
   (e) Give you practical advice for English learning (43.4%)
5. Our current chatbot design contains a button that reveals the transcript of the chatbot utterance when clicked; should we keep this button?
   (a) **Yes! (87.9%)**
   (b) No, the transcript should be displayed directly and automatically (12.1%)
6. Do you need Mandarin translations of chatbot utterances?
   (a) **I only need translations for chatbot feedback (38.2%)**

(b) I need translations for everything that the chatbot says (28.5%)

(c) I don't need any translation (33.3%)

There are some additional free-form responses provided by the internet users filling out our survey. We intend to perform further analyses of the survey and publicly share the results to provide research directions for others in the field.

C Grammar Model Training Details

We use a train-validation split of 0.9-0.1 when training our models. Both models were fine-tuned using PEFT on a single GPU for 10 epochs. The Llama-2 7B model was trained with an initial learning rate of `2e-4` and a batch size of 4. The Flan-T5-XL grammar model was trained with default parameters. The best checkpoints according to evaluation losses were selected.

D Conversation Model Training Details

For the conversation data, we use a train-validation split of 0.95-0.05. The Llama-2 model was PEFT-trained on a single GPU for 10 epochs, with an initial learning rate of `2e-4` and a batch size of 4. The best checkpoint according to evaluation losses was selected.

E Grammar Correction Hierarchy

We reference an online resource for grammatical error hierarchy[4] to establish the hierarchy of grammar errors recognized by our system.

In this hierarchy, errors are divided into tiers based on severity. Different error tiers correspond to different tolerance levels. For instance, if an error is tier #1 with a tolerance level 1, the error is immediately corrected (grammatical feedback is given on this error); if an error is tier #3 with a tolerance level 5, then this error will only be corrected if the user has made the same type of error for five times in one conversation. We detail this hierarchy in Table 7.

F Conversation Data Synthesis

F.1 Topics

Table 8 details the number of topics per broad topic area. For a complete list of topics, please see Appendix M.

[4] https://bcourses.berkeley.edu/courses/1196299/files/66663155/download?wrap=1

Tier	Errors	Tol.
#1	Word Order, Wrong Verb Tense, Incorrect Verb Form, Incorrect Preposition, Missing Preposition, Unnecessary Preposition, Wrong Collocation	1
#2	Subject-Verb Disagreement, Incorrect Singular/Plural Noun Agreement, Incorrect Possessive Noun, Incorrect Determiner	3
#3	Incorrect Auxiliary Verb, Incorrect Part of Speech, Missing Word Related To Verb Form, Missing Word Related To Verb Tense, Missing Determiner, Missing Verb, Missing Adjective, Missing Adverb, Missing Auxiliary Verb, Missing Adpositional Phrase, Missing Conjunction, Missing Particle, Missing Noun, Missing Pronoun, Unnecessary Determiner, Unnecessary Verb, Unnecessary Word Related To Verb Form, Unnecessary Word Related To Verb Tense, Unnecessary Adpositional Phrase, Unnecessary Adjective, Unnecessary Adverb, Unnecessary Auxiliary Verb, Unnecessary Conjunction, Unnecessary Particle, Unnecessary Noun, Unnecessary Pronoun, Spelling Error	5

Table 7: The grammar error hierarchy that we employ in our system.

F.2 Generation Diversity Issues and Corresponding Prompt Adjustments

In Li et al. (2023), ChatGPT is first prompted to generate two distinct personas, one generic persona (Person 1, often assumed to be American by ChatGPT), and one Chinese college student persona (Person 2). The LLM is then asked to generate a conversation using these personas, where Person 1 should lead the conversation by asking questions and sharing engaging anecdotes when appropriate. Multiple conversations with various persona choices are generated for each textbook topic and

Topic Area	Topic Counts
Food	36
Books	43
Movies	44
TV shows	31
Music	45
Hobbies	34
English learning	10
Total	243

Table 8: Number of topics per topic area.

corresponding vocabulary set.

We first adjust the requirements of the conversation generation step to make Person 1 more empathetic and attentive to Person 2. We also remove the vocabulary constraint and request that the generated conversations be spoken. However, we notice an alarming homogeneity in the generated conversations in terms of Person 2's preferences. For example, out of seven conversations generated about favorite foods, three feature hot pot, three feature dumplings, and one features Peking Duck. For favorite songs, "The Moon Represents My Heart" is Person 2's favorite in six out of ten conversations, and generic old Chinese songs are favorites in the other conversations. These overly repetitive examples do not represent the general population of Chinese college students. As a result, to enhance the diversity of our training data, we define Person 2 as someone whose first language is not English.

F.3 Data Synthesis Prompts

The following is the prompt used to generate different personas:

```
personas_prompt = ("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 but with
                different information; Person
                2's native language is not
                English.")
```

After generating the personas, given a specific topic, we use the following prompt to generate 10 conversations about this topic within the same ChatGPT prompting session:

```
convo_prompt = (
    'Generate a single spoken conversation
        between these two people as Person 1 and
         Person 2 about the topic "{topic}".\n'
    "Please take into account their distinct
        personalities and their backgrounds.
        Begin the conversation with Person 1.\n"
    "Person 1 should guide the conversation by
        asking more questions; Person 1 should
        also be attentive to Person 2's
        interests and ask Person 2 to say more.\
        n"
    "Person 1 should be able to make specific
        recommendations to Person 2 if requested.
         Person 2 should feel free to ask for
        recommendations from Person 1 if
        appropriate.\n"
    "Begin the conversation with Person 1. Person
         1 does not know any information about
        Person 2 unless Person 2 brings it up.
        Person 1 should not recommend
        restaurants, stores, or recipes. Keep
        utterances colloquial. Person 1 should
        discuss the recommendation directly in
        conversation, rather than saying they
        will send the recommendations later. The
         conversation should last at least 10
        turns.")
```

F.4 Data Filtering Mechanism

We apply the same data formatting filtering as Li et al. (2023) (e.g. making sure the conversation starts with Person 1, etc). We additionally include filtering mechanisms for our specific issues in dialogue generation. Specifically:

1. Person 1 would make assumptions about Person 2 without Person 2 mentioning it (e.g. asking Person 2 whether they miss Brazil even though Person 2 has not mentioned that they are Brazilian). This is likely because ChatGPT assumes Person 1 knows Person 2's persona.

2. Person 1 would offer to send their recommendations via private message. While this is likely in everyday conversation, since our chatbot does not have a mechanism for private messages, this is considered a failure mode.

We now present our filtering mechanisms. Given a dialogue history string, we provide ChatGPT with the following prompt:

```
"Does Person 1 in the following conversation
    make assumptions about Person 2 without the
    user bringing it up first? Answer with yes
    or no.\n\n" + dialogue_string
```

If the ChatGPT response starts with "yes", then we filter out this dialogue.

If not, we pass the conversation through one more layer of filtering:

```
"Does Person 1 in the following conversation
    make specific recommendations when requested
    ? If Person 2 does not request specific
    recommendations, answer \"Yes\". Answer with
     yes or no.\n\n" + dialogue_string
```

If the ChatGPT response starts with "no", then this conversation is pruned.

F.5 Topic Distribution over Conversations

Topic	Conversation Counts
Food	124
Books	243
Movies	209
TV shows	167
Music	233
Hobbies	195
English learning	56
Total	1227

Table 9: Number of conversations per topic after pruning

G Personalization Details

G.1 Short and Succinct Feedback

When the participant selects that they prefer succinct feedback and their experimental condition allows for feedback personalization, we use the following prompt along with the past three student utterances (`convo`) and the original generated feedback (`output`):

```
Given the following utterances by a student
    learning English as the context:\n\n{
    convo}\n\nAnd a piece of feedback:\n\n{
    output}\n\nMake it more succinct and
    concise while retaining the original
    examples with their full context. Make
    the feedback colloquial and succinct.
    Don't use the word \"basic\". Try to
    shorten to at most 3 sentences.
```

G.2 Detailed Feedback with Examples

When the participant prefers their feedback to have more detail, we use the following sequence of prompts with the past three student utterances (`convo`) and the original generated feedback (`output`). We use a sequence of prompts instead of a single prompt because we notice that ChatGPT often overgenerates on the detail and makes the feedback too long.

```
Given the following utterances by a student
    learning English as the context:\n\n{
    convo}\n\nAnd a piece of feedback:\n\n{
    output}\n\nCreate a new piece of
    feedback with more context-specific
    examples supporting the feedback. Make
    the feedback colloquial, as if spoken in
     conversation. Don't use the word \"
    basic\".
```

Upon obtaining the ChatGPT generation, we use the following prompt to shorten the feedback:

```
Shorten your response to 3 - 4 sentences
    while retaining necessary information
    and detail.
```

H Transition Improvements

H.1 Overview

After receiving a piece of feedback, the user may have some questions about the feedback that are directly related to English learning. If that is the case, our conversation model may not be able to handle them well since they are better trained on open-domain chit-chat. Therefore, we should ensure to use ChatGPT to handle these queries.

We classify whether a user utterance is a relevant query using ChatGPT. If the query is directly related to English learning, ChatGPT is instructed to produce a response. This response is presented to the user directly without any modification. If the user utterance is not a query or the query is not directly related to English learning, the ChatGPT response is processed such that all questions are removed from the response. This altered response is then concatenated with a connector phrase with a recap of the conversation before the feedback, as well as the response from the conversation model prior to activating the feedback module.

H.2 Query Classification

Given a conversation history of the last three turns, we classify the query using the following ChatGPT prompt:

```
Given the following user-chatbot exchange:\n\
    n{convo_history}\n\nIs the latest user
    utterance asking for clarifications or
    English learning advice? Answer with yes
     or no.
```

H.3 ChatGPT Response Processing

We generate the response to the user utterance using this ChatGPT prompt:

```
Respond to the last user utterance as the
    Assistant based on the conversation
    context. Be colloquial and helpful. You
    only know English and Mandarin.
```

If the latest user utterance is not a query or a relevant query, we process the ChatGPT response to remove any questions. Specifically, we first tokenize the response into individual sentences and concatenate sentences that do not end with question marks together. We then use ChatGPT to very briefly summarize the pre-feedback conversation using this prompt:

```
Given the following conversation history:\n\n
    {convo}\n\nDescribe the current general
    topic with ONE SHORT PHRASE.
```

We then create a connector sentence employing a pre-defined set of connector phrases. `curr_topic` here refers to the conversation summary phrase.

```
f"Alright, let's continue our conversation
    about {curr_topic}.", f"Let's get back
    to our chat on {curr_topic}!",
f"Okay let's go back to our conversation
    about {curr_topic}.", f"Now back to our
    conversation with respect to {curr_topic
    }.",
f"Lets' go back to our chat. We just talked
    about {curr_topic}.", f"Let's keep
    chatting about {curr_topic}.",
"Okay, let's keep chatting.", "Let's go back
    to our conversation!", "Let's continue
    our chat!"
```

I Fixed Empathetic Responses

The full list of fixed empathetic response utterances is as follows:

1. I understand that learning English can be a difficult process, but you are doing great! Keep it up!
2. I have been seeing steady progress from you. English learning can be challenging, but you are doing so well!
3. I'm really proud of the progress you're making. Keep powering through it and you will see even more significant improvement in your spoken English!
4. Your hard work on learning English is really paying off. You have been nailing this conversation thus far!
5. I can see how much you've improved in your spoken English. Keep it up and you will see even more improvement in no time!
6. Keep up the great work in practicing your spoken English; every effort brings you a step closer to fluency.

J Further Experiment Details

J.1 Informed Consent

We present our informed consent form through this link: https://github.com/siyan-sylvia-li/EDEN/blob/main/experimental_data/Informed%20Consent%20-%20Public%20-%20Google%20Forms.pdf.

J.2 Pre- and Post-surveys

In the pre-survey, in addition to the L2 grit survey, the participants are asked to complete the following questions about their English proficiency:

1. How many years have you been learning English?
2. How is your English proficiency? (1 is beginner and 5 is proficient)
3. What was your IELTS score?
4. What was your TOEFL score?
5. How much English do you speak every day? (1 is Mandarin only, 5 is English only)

In the post-survey, the participants are first asked to complete the questions about general conversation quality, then the questions about chatbot PAS, and then, finally, the L2 grit survey.

K Example Conversations

In the following examples, we use Red texts to indicate adaptive empathetic feedback, Blue texts to indicate fixed empathetic feedback, and Orange texts to indicate grammatical feedback.

K.1 Adaptive Empathetic Feedback Dialogue Example

USER: No, I Sorry, I normally just just How should I describe Normally when I cooking I just follow in my mind and you know I'm from China. I know the recipe taste like and I know what kind of the sauce I should put in when I cooking
BOT: You're doing well with the topic, but work on your grammar to sound smoother. For instance, say "I know what the recipe should taste like" instead of "I know the recipe taste like." Keep practicing to get even better! Does that sound alright to you?
USER: Yes, thank you. Thank you for the instruction. And we move on to the next topic.
BOT: Great! If you're ready, let's jump into the next topic. Just let me know if you need any help

or have any questions along the way. Okay let's go back to our conversation about recipe sources. Ah, that makes a lot of sense. It's wonderful that you have such a strong connection to your culture and food. It's like a bridge to your roots, right? I can imagine that cooking must bring back a lot of memories and emotions. Have you ever tried to share some of your Chinese dishes with friends or family here?
USER: Yeah, definitely. When we have a party with my friends, Chinese friends, yeah. Like I said before, we're cooking together and enjoy the meal together. We even can discuss which one is authentic, which one... What's that preference?
BOT: It's great that you're so thoughtful about the authenticity. It's important to have those discussions, especially when sharing something so meaningful. I think it's a beautiful way to connect with others. Speaking of which, I recently had a similar experience with a Peruvian dish. It was interesting to see how others perceived it and how it was passed down.
USER: Yeah, that can make a topic when we hang out together. So people can share in the life experience and we can have more topics, more subjects to talk.
BOT: Maybe you meant "So people can share in their life experiences and we can have more topics and subjects to talk about.". Does that sound alright to you?

K.2 Fixed Empathetic Feedback Dialogue Example

USER: Recently I'm watching Friends, that's a really classic sitcom but I'm really enjoying that. Now I'm coming to the sixth episode. So do you like this sitcoms and how do you find the characters?
BOT: Friends is a great choice! I've watched it a few times, and I can see why it's so popular. The characters are definitely unique and relatable in their own ways. How do you find the comedy in the show, and do you have a favorite character?
USER: That's an interesting question. Before I liked Ross because I think he's really reliable and I like smart people but now I gradually like Chandler more because I found he's really you know the type of dream partner I would say. He treats Monica really well and he's funny by the way yeah I like humorous people
BOT: I have been seeing steady progress from you. English learning can be challenging, but you are doing so well!
USER: Thank you. Actually, I still think I lack competence and also the fluency in my speaking English. Oftentimes, I don't know how to express my feeling quite well and also it's difficult to understand the idioms, some common expressions. So do you have any suggestions according to that, like how to improve my daily English speaking?
BOT: No worries at all! Improving your English speaking skills is totally doable. One suggestion is to practice speaking with someone regularly, maybe a classmate or a language exchange partner. You could also try watching English movies with subtitles to pick up on common expressions and idioms. Remember, practice makes perfect, so keep at it! Good luck!
USER: Thank you for saying that. Maybe I want to improve my pronunciation because I often feel my speaking is not authentic as a native speaker and you can easily tell that I'm a foreigner. I really want to improve that part. Do you have any suggestions?
BOT: No worries at all! Improving your pronunciation is a great goal. One tip is to practice speaking slowly and focus on the sounds of each word. You could also try mimicking native speakers or even recording yourself to hear how you sound. Keep at it, and don't be too hard on yourself - progress takes time! Good luck! Let's go back to our conversation! I can see why Chandler's character appeals to you. His wit and humor really make the show more enjoyable. It's great to hear that you're enjoying the show so far. Do you think the cultural differences between the characters add to the humor or do you find that they sometimes create challenges in understanding their dynamics?

K.3 Grammatical Feedback Dialogue Example

USER: For that part, I do agree it has brought positive impact onto the whole entertainment business with more representations of Asian experiences.
BOT: I believe you wanted to say "For that part, I do agree it has brought a positive impact onto the whole entertainment business by representing more Asian experiences." Does that sound good?
USER: Oh yeah, that sounds good.

L Additional Details for Results

L.1 PAS and Conversation Quality

We analyze the relationship between overall PAS and our different measures for conversation quality,

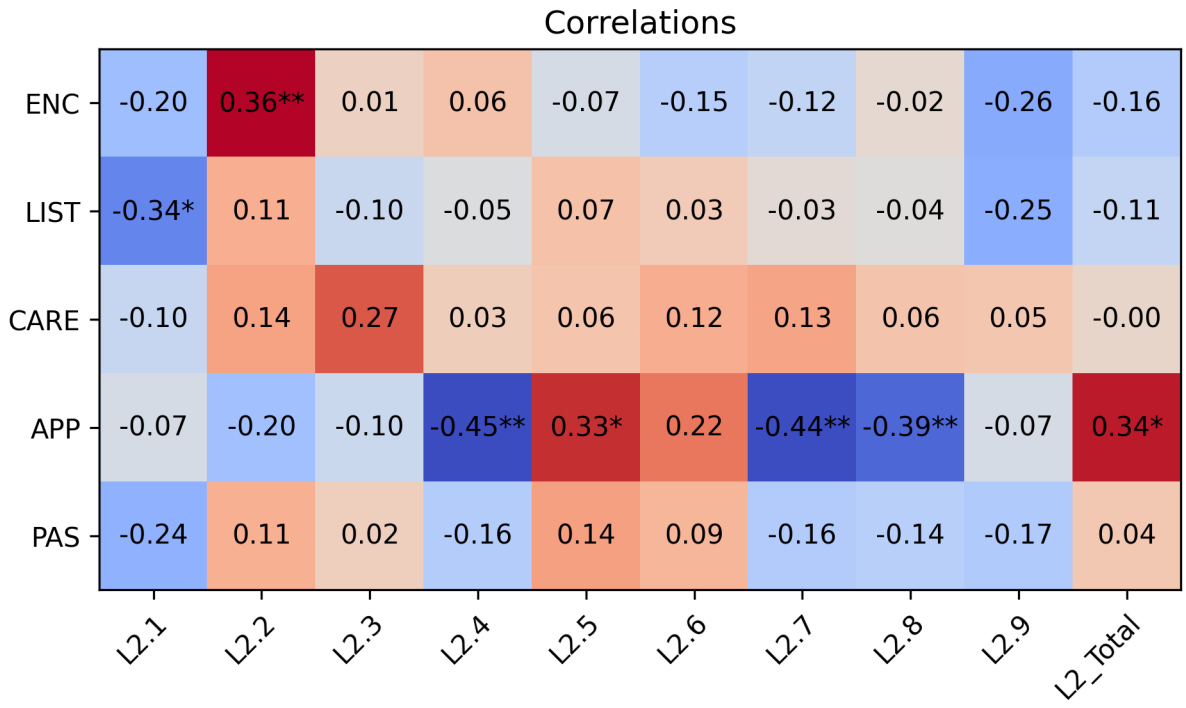

Figure 4: Correlations between different measures for chatbot PAS and L2 grit. **= $p < 0.05$, *= $p < 0.1$

as well as the relationship between overall L2 grit changes and our measures for conversation quality. We present the result of our correlational analysis in Table 10.

Measure		QUAL	CONF	USE
PAS	Coef.	0.28	0.54	0.62
	p-value	0.13	0.0018	0.0002
ΔL2_Total	Coef.	-0.17	-0.19	0.05
	p-value	0.35	0.31	0.78

Table 10: Correlations between PAS and different measures for conversational quality, as well as correlations between L2 grit and measures for conversational quality.

L.2 PAS and L2 Grit

Here, we analyze the relationship between PAS measures and L2 grit, rather than changes in L2 grit. We present the resulting correlation matrix in Figure 4. We see that APP has negative correlations with the reverse-coded items in the L2 grit scale, and is positively correlated with total L2 grit in the post-survey. We see that LIST also has a weak negative correlation with L2.1, similar to how CARE correlates negatively to ΔL2.1.

M Complete Topics List

```
FOOD
['Cooking traditions in family gatherings', '
    Exploring cultural significance through food
     memories', 'Nostalgic meals from childhood
    ', 'Evolution of taste preferences over time
    ', 'Food-related rituals and celebrations',
    'Culinary adventures while traveling', '
    Impact of favorite food-related memories on
    overall well-being', 'Favorite food', '
    Cultural significance of favorite foods', '
    Psychological aspects of comfort foods', '
    Historical origins of popular dishes', '
    Regional variations in favorite foods', '
    Impact of advertising on food choices', '
    Fusion cuisine and blending of flavors', '
    Favorite cuisine', 'Fusion cuisines
    incorporating favorite elements', 'Health
    benefits of favorite cuisines', 'Popular
    street foods within favorite cuisines', '
    Vegan/vegetarian adaptations of favorite
    cuisines', 'Cultural significance of
    ingredients in favorite cuisines', 'Cooking
    techniques specific to favorite cuisines', '
    Famous chefs and restaurants specializing in
     favorite cuisines', 'Street food
    preferences', 'Global street food culture',
    'Health considerations in street food', '
    Popular street food vendors around the world
    ', 'Street food festivals and events', 'DIY
    street food recipes', 'Historical evolution
    of street food', 'Street food and cultural
    identity', 'Sustainable practices in street
    food markets', 'Street food safety
    regulations', 'Street food fusion trends', '
    The Role of Food in Celebrations and
    Festivals', 'Cultural significance of
```

```
    traditional dishes in festivals', 'Evolution
     of festival foods over time']

HOBBIES
['Finding time for hobbies', 'Time management
    techniques', 'Exploring leisure activities',
     'Prioritizing personal interests', '
    Balancing work and leisure', 'Creating a
    hobby schedule', 'Discovering passion
    projects', 'Incorporating relaxation into
    daily routine', 'Maximizing free time', '
    Setting goals for hobbies', 'Joining hobby
    groups or clubs', 'What to do during free
    time', 'Hobbies to Pursue', 'Outdoor
    Activities to Try', 'Creative Projects to
    Start', 'Indoor Activities for Relaxation',
    'DIY Projects to Explore', 'Social
    Activities to Engage In', 'Learning New
    Skills', 'Volunteering Opportunities', '
    Cultural Events to Attend', 'Wellness
    Practices for Self-care', 'New hobbies', '
    Picking up new hobbies', 'Outdoor activities
    ', 'Crafting and DIY projects', 'Gardening
    and urban farming', 'Cooking and baking', '
    Fitness and exercise routines', 'Music
    production and learning instruments', '
    Painting and drawing', 'Photography and
    videography', 'Creative writing and
    journaling', 'Board games and tabletop
    gaming']

MOVIES
['Favorite movie', 'Movie genres and their
    characteristics', 'Impact of favorite movies
     on personal taste', 'Analysis of favorite
    movie soundtracks', 'Cultural significance
    of favorite movies', 'Evolution of movie
    preferences over time', 'Favorite movie
    directors and their filmography', '
    Psychology behind attachment to favorite
    movies', 'Societal influence on favorite
    movie choices', 'Comparing favorite movies
    with critical acclaim', 'The role of
    nostalgia in favorite movie selection', '
    Favorite movie director', 'Filmography
    analysis of favorite movie director', '
    Influence of favorite movie director on
    modern cinema', 'Cinematic style of favorite
     movie director', 'Collaborations with
    actors/actresses by favorite movie director
    ', "Favorite movie director's impact on the
    industry", "Favorite movie director's
    signature themes and motifs", "Evolution of
    favorite movie director's directing
    techniques", "Comparison of favorite movie
    director's works with contemporaries", "
    Behind-the-scenes insights into favorite
    movie director's creative process", 'Legacy
    of favorite movie director in film history',
     'Favorite movie genre', 'Action-packed
    films', 'Romantic comedies', 'Sci-fi and
    fantasy flicks', 'Horror movies', '
    Historical dramas', 'Animated features', '
    Mystery and thriller genres', 'Documentaries
    ', 'Musical films', 'Adventure movies', '
    What makes a good movie', 'Character
    development in films', 'Plot structure and
    storytelling techniques', 'Visual aesthetics
     and cinematography', 'Soundtrack and
    musical score impact', 'Effective use of
    symbolism and motifs', 'Genre conventions
    and audience expectations', 'Impact of
    pacing and editing on viewer engagement', '
    Dialogue and scriptwriting excellence', '
    Cultural and societal influences on film
    reception', 'Directorial style and vision
    manifestation']

MUSIC
['Favorite song', 'Music genres', 'Lyric
    analysis', 'Musical composition techniques',
     'Influence of culture on music preferences
    ', 'Evolution of music over decades', '
    Impact of technology on music production', '
    Music therapy benefits', 'Famous songwriters
     and their work', 'Music and emotions', '
    Role of music in society', 'Favorite musical
     artist', 'Favorite band', 'Favorite musical
     genre', 'History of jazz music', 'Evolution
     of rock and roll', 'Impact of hip hop
    culture', 'Classical music composers', 'Folk
     music traditions around the world', '
    Influence of electronic music on modern
    culture', 'Pop music trends and analysis', '
    Traditional music instruments of various
    cultures', 'Fusion genres in contemporary
    music', 'Music therapy and its benefits', '
    How music makes you feel', 'Psychological
    effects of music', 'Emotional impact of
    music', 'Music therapy benefits', '
    Neuroscience of music and emotions', 'Music
    and mood regulation', 'Cultural influences
    on music perception', 'Music and memory
    recall', 'Physiological responses to music',
     'Music and stress reduction', 'Social
    bonding through music', 'Playing musical
    instruments', 'Music theory', 'Learning
    techniques', 'Instrument maintenance', '
    Historical development of instruments', '
    Musical genres', 'Famous musicians', 'Music
    composition', 'Instrument accessories', '
    Performance techniques']

TV SHOWS
['Favorite TV Show', 'Character Development in
    TV Shows', 'Impact of TV Shows on Culture',
    'Evolution of TV Show Genres', '
    Representation in Television', 'Exploring TV
     Show Soundtracks', 'The Role of Television
    in Storytelling', 'Favorite TV character', '
    Character development in TV shows', 'Impact
    of TV characters on audience', 'Evolution of
     TV show protagonists', 'Analysis of popular
     TV show archetypes', 'Gender representation
     in TV show characters', 'Cultural
    significance of iconic TV characters', '
    Character arcs in long-running TV series', '
    Favorite TV genre', 'Favorite Comedy Series
    ', 'Favorite Drama Series', 'Favorite Crime
    Shows', 'Favorite Science Fiction Series', '
    Favorite Fantasy Series', 'Favorite
    Documentary Series', 'Favorite Reality TV',
    'Favorite Animated Series', 'Favorite
    Historical Drama', 'Favorite Thriller Series
    ', 'TV show binge-watching habits', '
    Streaming platforms usage', 'Effects of
    binge-watching on sleep', 'Psychological
    impact of binge-watching', 'TV show reboots
    and revivals']
```

BOOKS
['Favorite book', 'Favorite novel', 'Favorite non-fiction', 'Favorite fiction', 'Favorite author', 'Favorite authors', 'Favorite book genres', 'Literary influences', 'Writing styles', 'Character development techniques', 'Plot structures', 'Narrative perspectives', 'Symbolism in literature', 'Authorial voice', 'Classic literature', 'Literary analysis techniques', 'Historical context in literature', 'Themes in classic literature', 'Famous authors of classic literature', 'Impact of classic literature on society', 'Gender roles in classic literature', 'Adaptations of classic literature in film and theater', 'Book clubs', 'Reading habits', 'Community engagement through books', 'Social impact of book clubs', 'Diversity in reading selections', 'Virtual book club trends', 'Must-read books', 'Classic literature books', 'Modern fiction books', 'Non-fiction bestsellers', 'Biographies and memoirs', 'Science fiction and fantasy books', 'Self-help and personal development books', 'History and politics books', 'Philosophy and spirituality books', 'Crime and mystery books', 'Young adult literature books', 'Bookstores and libraries', 'Audiobooks vs. physical books', 'Book adaptations (movies, TV shows, etc.)', 'Fiction vs. non-fiction']

ENGLISH LEARNING
['Vocabulary acquisition and expansion', 'Grammar rules and structures', 'Pronunciation practice', 'Reading comprehension strategies', 'Writing skills development', 'Listening comprehension exercises', 'Speaking fluency and conversation practice', 'Idioms and expressions', 'Cultural aspects and context in English language learning', 'Test preparation (e.g., TOEFL, IELTS, Cambridge exams)']