A Conversational Intelligent Tutoring System for Improving English Proficiency of Non-Native Speakers via Debriefing of Online Meeting Transcriptions

Juan Antonio Pérez-Ortiz[°*], **Miquel Esplà-Gomis**[°], **Víctor M. Sánchez-Cartagena**[°], **Felipe Sánchez-Martínez**[°], **Roman Chernysh**[°], **Gabriel Mora-Rodríguez**[°], **Lev Berezhnoy**[°]

[°] Universitat d'Alacant, Spain {japerez,mespla@ua.es}
[*]Valencian Graduate School and Research Network of Artificial Intelligence, ValgrAI, Spain

Abstract

This paper presents work-in-progress on developing a conversational tutoring system designed to enhance non-native English speakers' language skills through post-meeting analysis of the transcriptions of video conferences in which they have participated. Following recent advances in chatbots and agents based on large language models (LLMs), our system leverages pre-trained LLMs within an ecosystem that integrates different techniques, including in-context learning, external non-parametric memory retrieval, efficient parameter fine-tuning, grammatical error correction models, and error-preserving speech synthesis and recognition. While the system is still in development, a preliminary pilot evaluation of a prototype has been conducted with L2 English students.

1 Introduction

In an increasingly interconnected world, the ability to communicate effectively in English has become a vital skill, especially in professional settings where English has firmly established itself as the *lingua franca* (Nickerson, 2005; Shegebayev, 2023). However, this requirement often leads to challenging situations for many non-native speakers who, when participating in meetings, presentations, and discussions conducted in English, frequently find themselves navigating the complexities of the language under the potential scrutiny of more fluent colleagues. This dynamic can create a stressful environment, hindering effective communication and the free flow of ideas, leading to misunderstandings, and impacting the confidence and performance of less-proficient speakers (Aichhorn and Puck, 2017). These linguistic shortcomings are often silently noted by other participants, but rarely addressed in a constructive manner, and the very settings where these individuals most frequently use English are not leveraged as opportunities for improvement.

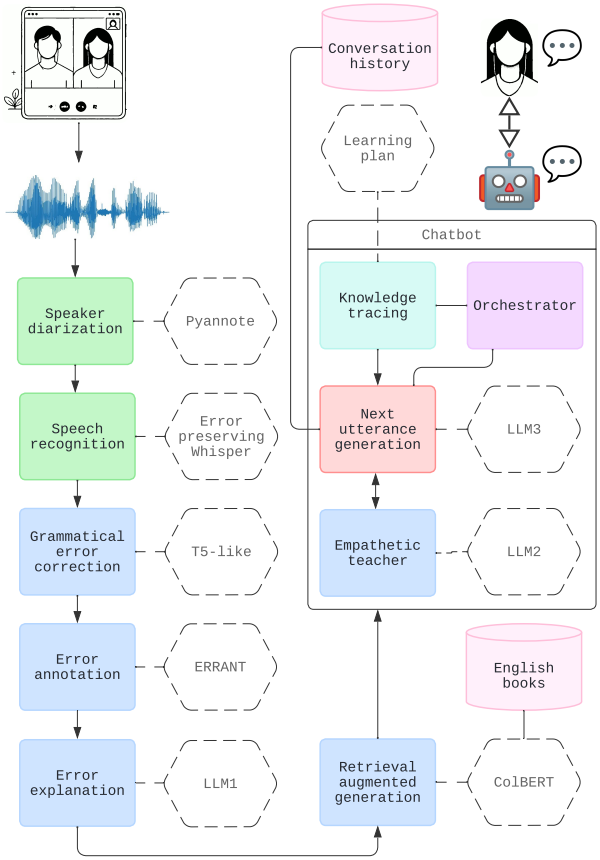

Figure 1: Main components of the DeMINT system. Section 3 describes each component thoroughly.

Although a human tutor could provide valuable feedback and guidance to help non-native speakers improve their language skills, this solution is often impractical due to logistical constraints, financial considerations, or the reluctance to introduce additional complexity into an already demanding professional life. To address this gap, we propose an automated language debriefing system that leverages the transcripts of online meetings to

Juan Antonio Pérez-Ortiz, Miquel Esplà-Gomis, Víctor M. Sánchez-Cartagena, Felipe Sánchez-Martínez, Roman Chernysh, Gabriel Mora-Rodríguez and Lev Berezhnoy. A Conversational Intelligent Tutoring System for Improving English Proficiency of Non-Native Speakers via Debriefing of Online Meeting Transcriptions. *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2024).* Linköping Electronic Conference Proceedings 211: 187–198.

provide feedback and guidance to non-native English speakers (L2-English), thus mimicking the role of a language instructor. Our system, called DeMINT after the project in which it was developed, is implemented as an educational chatbot (Du and Daniel, 2024) that interacts with users in a conversational manner, thereby transforming everyday professional interactions into valuable opportunities for language improvement.

Conversational intelligent tutoring systems (ITS) are set to revolutionize the field of education, offering one-to-one, personalized, interactive, engaging, and inclusive learning experiences to students. Their application in computer-aided language learning (CALL) is particularly promising, as contemporary large language models (LLMs) show remarkable capabilities in language understanding and generation. While such systems were explored in the past (Jia, 2009; Bibauw et al., 2019), only with the advent of contemporary LLMs have functional implementations become feasible.

Our ITS aims to leverage LLMs in a CALL application to improve speakers' language skills through interactive, personalized, and error-driven conversations. A functional prototype has been evaluated in a pilot study with L1 Spanish/L2 English learners. The source code, along with links to models and datasets, is available online.[1]

The rest of the paper is organized as follows. Section 2 reviews related work on chatbots in education. After that, Section 3 describes DeMINT's design and its main components. Then, Section 4 outlines the human evaluation. After the conclusions and potential future work described in Section 5, the ethical considerations of the project and the main limitations are highlighted.

2 Related Work

The year 2022 marks a turning point where the capabilities of LLMs for conversational and educational tasks, in general, and ITS, in particular, became evident. Despite this, prior research had already demonstrated the potential benefits of using traditional chatbots within dialog-based CALL scenarios for L2 learners (Jia, 2009; Bibauw et al., 2019; Huang et al., 2022; Yang et al., 2022), identifying pedagogical, technological, and social affordances (Jeon, 2024).

Shortly after the release of ChatGPT in November 2022, several studies explored its potential for L2 teaching. A survey among English-as-foreign-language faculty instructors by Mohamed (2024) highlighted ChatGPT's ability to enhance proficiency and motivation, while also emphasizing the need to address limitations and ethical concerns. A meta-analysis by Zhang et al. (2023) of 18 articles on chatbot-assisted language learning concluded that "using chatbots for language learning has a positive impact, and the learning outcomes are better than those in non-CALL situations." A more recent meta-study by Cislowska and Acuña (2024) observed that "the use of chatbots can positively affect students' attitudes toward learning a foreign language, enhancing motivation, interest, fun, proactivity, and learning commitment"; however, they also noted that "the novelty effect may decrease motivation over time, and lacking a human factor may fail to meet emotional needs and decrease motivation." Several other recent reviews have reached similar conclusions (Labadze et al., 2024; Du and Daniel, 2024).

Additionally, the emergence of commercial AI-driven language learning assistants developed by companies like Duolingo,[2] Google,[3] or TalkPal[4] underscores the growing importance and effectiveness of LLM-based CALL systems. In spite of the potential of these systems, we are not aware of many open-source projects that implement a comprehensive conversational ITS for L2 learning as ours, especially one that leverages the transcripts of online meetings to provide feedback and guidance to L2-English speakers.

3 System Description

Our system design draws from recent chatbots like BlenderBot3 (Shuster et al., 2022) which are built as a pipeline of different modules that mainly consist of LLMs fine-tuned for specific tasks.[5]

A diagram of the main components of DeMINT is shown in Figure 1 on the first page. As can be seen, the system is composed of several modules that interact with each other to provide a compre-

[1] https://github.com/transducens/demint

[2] https://blog.duolingo.com/duolingo-max

[3] https://research.google/blog/english-learners-can-now-practice-speaking-on-search

[4] https://talkpal.ai

[5] This also resonates, albeit on a smaller scale, with the revitalization of Minsky's societies of mind theory (Minsky, 1986) in the form of natural language-based societies of LLMs and other machine learning models mindstorming together to solve a problem (Zhuge et al., 2023).

hensive tutoring experience. Some of them are based on pre-trained LLMs, pre-trained sequence-to-sequence models or ad-hoc models, while others rely on external resources such as textbooks on English grammar. Next sections describe each of these modules in detail. The pipeline of modules outside of the chatbot box in Figure 1 is run offline before the chatbot starts interacting with the user.

3.1 Diarization

The pipeline starts by processing the audio recordings of the target online meeting and identifying the segments corresponding to each speaker. This is done by using the library `pyannote.audio` (Bredin, 2023; Plaquet and Bredin, 2023), which relies on a neural speaker diarization model (Takashima et al., 2021). The diarization process returns the start and end times of each speaker turn, as well as the speaker ID.

The audio fragments corresponding to each speaker are then individually processed by the speech recognition system described in the next section. Remarkably, an alternative approach has been recently proposed where diarization and transcription are performed in parallel, and the outputs are subsequently combined.[6]

3.2 Speech Recognition

As our error analysis pipeline is performed on the written transcriptions of the online meetings, a speech-to-text (STT) model is needed to transcribe the utterances for each speaker. Our initial approach was to directly use open-weight pre-trained models such as Whisper (Radford et al., 2023), but preliminary tests showed that they were not entirely suitable for our purposes, due to the fact that their strong internal language model tends to correct some of the grammatical errors in the utterances. For example, the Whisper model would often transcribe "I *doesn't know" as "I don't know", which is unacceptable for our purposes as the original grammar errors need to be faithfully preserved. Consequently, our system includes a custom error-preserving STT model that retains more grammatical errors. This model is obtained by fine-tuning Whisper on a custom dataset of spoken sentences with grammatical errors that we specifically created for our system.[7]

The ad-hoc dataset comprises both synthetic and natural texts containing grammatical errors. The natural texts are sourced from the COREFL dataset (Lozano et al., 2020), which contains essays by non-native students with varying levels of English proficiency.[8] COREFL includes some audio recordings of students reading their texts, as well as written compositions. However, since only a small percentage of the texts have corresponding audio recordings, we have also converted written texts into audio using the StyleTTS2 text-to-speech (TTS) model (Li et al., 2023), which allows us to synthesize each text with multiple voices, thereby increasing the diversity of the training data. On the other hand, the synthetic texts come from the $\mathrm{C4}_{200\mathrm{M}}$ dataset (Stahlberg and Kumar, 2021), which contains heterogeneous grammatically incorrect sentences synthetically generated via a corruption model.[9] We have converted these sentences into audio using the same StyleTTS2 model. The resulting dataset contains 32,000 speech training samples, 1,000 validation samples, and 1,000 test samples. The training set is composed of 28,592 utterances from $\mathrm{C4}_{200\mathrm{M}}$, 814 audios directly obtained from COREFL, and 2,594 synthetic utterances generated from the COREFL written texts. The test and validation sets are similarly divided between the two sources. This dataset is then used to fine-tune Whisper, which is subsequently employed to transcribe the audio from the online meetings. Further details on the hyperparameters used for model fine-tuning can be found in the appendix. The two resulting models—one based on the original Whisper model and the other on its distilled version, which is the one we ultimately used—are available on the HuggingFace hub.[10][11]

[6] https://huggingface.co/blog/asr-diarization

[7] Michot et al. (2024) recently demonstrated that certain CTC-based encoder models corrected slightly fewer errors than the encoder-decoder-based Whisper model, which they attributed to the reduced influence of the language model, but this came at the expense of degraded overall performance. As a result, we continue to use Whisper in our system. It is worth noting that their study addresses a similar challenge, aiming to develop error-preserving STT models. While our approach is primarily automatic, their work involves the collection and annotation of a corpus containing English grammatical errors from young learners.

[8] Given that our evaluation will primarily involve students whose mother tongue is Spanish, we use only the subset of COREFL produced by Spanish students.

[9] In order to avoid the fine-tuned model relying too much on ungrammatical utterances, we add clean utterances from the *correct side* of $\mathrm{C4}_{200\mathrm{M}}$ to the training dataset as well.

[10] https://huggingface.co/Transducens/error-preserving-whisper

[11] https://huggingface.co/Transducens/error-preserving-whisper-distilled

Both datasets complement as $C4_{200M}$ provides a wide range of sentences and errors, although with a limited repertoire of voices, while COREFL offers a more diverse set of voices, accents, and natural errors. The COREFL dataset has the additional advantage of allowing our system to adapt to the accents and errors typically made by L1-Spanish speakers, who are the users in our pilot study. Due to the licensing restrictions of COREFL, only the dataset portion based on the $C4_{200M}$ dataset has been released on the HuggingFace hub as the Synthesized English Speech with Grammatical Errors Dataset (SESGE).[12]

Each transcribed utterance is split into sentences[13] before proceeding to the next step, and each sentence is associated with the speaker ID.

3.3 Grammatical Error Correction

Core to our work, *grammatical error correction* (GEC) is a well-known NLP task that aims to correct grammatical errors in a given text (Bryant et al., 2023; Omelianchuk et al., 2024). There are established shared tasks (Bryant et al., 2019) and datasets such as FCE (Yannakoudakis et al., 2011), NUCLE (Dahlmeier et al., 2013), Lang-8 (Mizumoto et al., 2011; Tajiri et al., 2012), W&I+LOCNESS (Bryant et al., 2019), and JFLEG (Napoles et al., 2017). A GEC model transforms a sentence with grammatical errors into a grammatically correct one.

Our system currently employs a relatively simple model obtained by fine-tuning the T5 encoder-decoder model (Raffel et al., 2020) on the JFLEG dataset,[14] but we are considering using more advanced state-of-the-art models such as GRECO (Qorib and Ng, 2023) or the ensembles provided by Omelianchuk et al. (2024).[15]

3.4 Error Annotation

Given the original and the corrected version of each sentence, we use the ERRANT toolkit[16] (Felice et al., 2016; Bryant et al., 2017) to extract and annotate the edits necessary to transform one sentence version into the other. ERRANT accomplishes this by applying an extended, linguistically-motivated version of the classical Levenshtein distance (Levenshtein, 1966), followed by a rule-based labeling of the edits. The resulting annotations are stored in the M2 format and then integrated into the JSON schema used as the intermediate format between the different components of our system.

3.5 Error Explanation

As ERRANT provides high-level annotations such as `R:VERB:SVA` (error in subject-verb agreement) without additional details, an LLM is used to generate finer-grained natural-language explanations of these errors via few-shot in-context learning. This aligns with recent works on using LLMs to further explain corrections made by GEC models (Fei et al., 2023; Kaneko and Okazaki, 2024; Song et al., 2024). These explanations will later *inspire* the chatbot's responses to the user.

Among all the open-weight, locally-installable LLMs available, we have found Llama-3.1-8B[17] to offer a good balance between speed and quality. Regarding the prompts used to query the model, we are considering using the DSPy framework (Khattab et al., 2023) to automatically generate them via DSPy's principled search mechanism (Khattab et al., 2022).[18]

3.6 Retrieval from Textbooks

Another component of the pre-processing pipeline is a module that retrieves information from English learning textbooks based on the errors being analyzed. This information will be one of the inputs provided to the chatbot's next-dialog-line generator at the end of the pipeline. We collected six PDF textbooks to be consulted under the retrieval-augmented generation (RAG, see below) approach, either open-licensed or available from `archive.org`. These English textbooks are valuable not only for their explanations of grammatical rules but also for the real examples of language usage they provide.[19]

[12] https://huggingface.co/datasets/Transducens/sesge

[13] Sentence splitting is achieved using the Python's package `sentence-splitter`.

[14] https://huggingface.co/vennify/t5-base-grammar-correction

[15] https://github.com/grammarly/pillars-of-gec

[16] https://github.com/chrisjbryant/errant

[17] https://huggingface.co/meta-llama/Meta-Llama-3.1-8B-Instruct

[18] We have found that DSPy's `compile` function is useful even for simple chains involving a single model, as it allows us to easily replace the model without manually rewriting the prompt, and also enforces a certain structure in the JSON output.

[19] We also plan to use an LLM as a source of this kind of grammatical information and examples, and to compare the results of both approaches.

Retrieval-augmented generation (RAG) encompasses a variety of techniques that integrate information from external documents into the generation process (Lewis et al., 2020). Naïve approaches to RAG involve segmenting documents into passages, computing an embedding for each passage, and storing both texts and embeddings in a vector database. Based on the current topic (each particular error in the use of English in our case), the most relevant passages are retrieved from the database by efficiently computing the similarity between an embedding of the topic and the embeddings of the passages. These selected passages are then provided to an LLM as a source of information for generating the output.

Our system employs the state-of-the-art ColBERTv2 model (Santhanam et al., 2022b), as implemented by the RAGatouille[20] library. ColBERTv2 computes token-level embeddings for passages and queries, making it more suited to our task than alternatives that compute a single dense embedding for each paragraph, as books' passages are likely to contain heterogeneous information such as grammar rules, open-domain examples, and exercises. ColBERTv2 is combined with a technique called *performance-optimized late interaction driver* (PLAID) (Santhanam et al., 2022a), which replaces conventional vector databases such as FAISS (Douze et al., 2024) with a more efficient and scalable approach based on using centroids of clusters of embeddings instead of the embeddings themselves. Additionally, the RAGatouille documentation states that its implementation of ColBERTv2 is robust in new domains and includes strong default settings, thereby eliminating the need for fine-tuning.

The above modules run offline prior to the debriefing session. Next, we describe those actively engaging in the chatbot's interaction with the user.

3.7 Empathetic Teacher

Another *ingredient* fed to the next-dialog-line generator comes from an LLM fine-tuned with real-life, ideally-empathetic teacher-student conversations. This model processes the recent conversation history and provides guidance on how a teacher might respond to the student's utterance. In order to obtain this model, we fine-tuned the Llama-3.1-8B model[21] with the following datasets: the Teacher-Student Chatroom Corpus, TSCCv2 (Caines et al., 2022), CIMA (Stasaski et al., 2020), the Multicultural Classroom Discourse Dataset (Rapanta et al., 2021), MathDial (Macina et al., 2023), and Conversational Uptake (Demszky et al., 2021). Some of the datasets were preprocessed in order to split very long conversations resulting in the figures shown in Table 1. The resulting collection of 6 503 conversation turns was split into 5 859 for training, 322 for validation, and 322 for testing, with each dataset contributing proportionally the same across these splits. Further details on the training hyperparameters are provided in the appendix. The fine-tuned teacher model is available on the HuggingFace hub.[22]

Dataset	Original		Split turns	
	Turns	Words	Turns	Words
TSCC v2	570	788k	1 074	786k
CIMA	1 135	44k	1 135	38k
MathDial	2 861	923k	2 876	879k
Multicultural	5	614k	643	614k
Uptake	774	35k	775	34k
Total	5 345	2 404k	6 503	2 351k

Table 1: Datasets used to train the empathetic teacher. Number of conversation turns and words in the original datasets and after splitting long conversations.

3.8 Orchestrator

The orchestrator is a simple Python program that iterates through the different errors and sentences to discuss them with the user during the debriefing session. For the current target error and sentence, the orchestrator prepares a complex prompt that includes the original sentence, the corrected sentence, the current error to review, the error annotation, the explanation of the error, the related information extracted from textbooks, the hints of the empathetic teacher's response, the short-term conversation history and a summary of the mid-to-long-term most relevant topics discussed with the user. Note that many of these items are by-products of the components described above. The orchestrator takes also into consideration the directives of the knowledge tracing module (see below) as regards the current state of the conversation flow and the user's understanding level. This long prompt will then be fed to the next-dialog-line generator.

[20] https://github.com/bclavie/RAGatouille

[21] https://huggingface.co/meta-llama/Meta-Llama-3.1-8B

[22] https://huggingface.co/Transducens/empathetic-teacher

The prompt consists of several key parts. First, the chatbot receives instructions on guiding the user through explanations, examples, and exercises to address the errors. It also outlines a set of user *intentions* for different interaction stages, with specific actions the chatbot must take for each. The prompt also includes the items generated in the pre-processing pipeline. Finally, concise instructions guide the chatbot to identify the user's intention, generate a suitable response, and output both following a JSON template. Additional guidelines ensure that responses are brief, engaging, and flexible beyond the provided data.

3.9 Knowledge Tracing

In order to guide the conversational flow, we first considered a traditional strategy based on a state transition model, with states representing the user's position in the learning path. However, we later found that the language model could manage the conversation flow autonomously via in-context learning and intention detection, without the need for extensive external intervention to track the dialog state. Transitions are therefore naturally handled by the LLM, based on the user's responses. Errors are prioritized based on frequency, according to the ERRANT's classes.

3.10 Next-Dialog-Line Generator

Although all LLMs used in the previously discussed components of DeMINT are implemented as local open-weight models, our preliminary experiments show that the best results are achieved when the next-dialog-line generator in particular is a more powerful LLM. Currently, GPT-4 accessed via the OpenAI API[23] is our preferred choice for this task.[24] This component faces the challenging task of generating the next line of the conversation based on the informative prompt prepared by the orchestrator. The output of this generator is then presented to the user as the chatbot's response.

3.11 Chatbot Interface

The interface is a simple web app built with `gradio`.[25] It shows the chatbot conversation in one column and the transcription, centered on the current sentence, in another. The user types their input, and the machine responds accordingly on the screen.

4 Human Evaluation

A preliminary evaluation[26] has been conducted through interactions between the chatbot and L1-Spanish/L2-English students. These students have been recruited through the Languages Service of our university, which maintains a pool of students registered for activities related to multilingualism promotion. This service retains information regarding the students' backgrounds, native languages, proficiency levels in languages, etc. Among the students willing to participate in this evaluation, 7 participants were selected, each dedicating approximately 10 hours to evaluation activities. We targeted students with B2/C1 levels of English according to the Common European Framework of Reference for Languages, and aimed to create a balanced group in terms of gender and diversity of backgrounds.

Fifteen video calls of approximately 10 minutes were organized among the selected students, with two or three participants per call. We employed role-playing games, specifically designed to engage students in English conversations. Role-playing games help avoid the difficulties associated with anonymizing real online meetings and allow us to control the topics and complexity of the conversations. Specifically, we have used the materials designed by Pitts (2015), which provide the context for the role-playing games, as well as preparatory questions to help students familiarize themselves with the topic. Students were given time to prepare for the online meeting. These video calls were recorded, and students then participated in a debriefing session with our chatbot to analyze errors in their use of English during the online meeting. Finally, students completed a survey to evaluate their interaction with the chatbot.

Feedback from the human evaluation addressed two main areas: overall user experience and the chatbot's effectiveness as an English tutor, with responses rated on a Likert scale from 1 to 5. Regarding the first aspect, participants were generally satisfied with the tool's performance and response time. In response to the question, *"Did you enjoy interacting with the chatbot?"*, all participants gave positive feedback, with a score of 4

[23] https://openai.com/api

[24] In particular, we use the `gpt-4o-2024-08-06` model, which, in addition to being one of the most powerful LLMs available today, includes the built-in feature *structured outputs* that enforces the generation of outputs in a specific JSON schema, thereby simplifying the ensuing parsing.

[25] https://github.com/gradio-app/gradio

[26] The empathetic teacher was disabled during evaluation.

or 5. However, fluency emerged as the system's main area requiring improvement, with an average score of 3. In terms of the chatbot's performance as an intelligent English tutor, the overall evaluation was positive, though some areas still require enhancement. The main concern of the participants in this evaluation was the accuracy of the chatbot in identifying speech errors, which received an average of 3. Other aspects, such as the chatbot's ability to understand their queries, or the usefulness of examples and resources provided by the chatbot, were rated with an average score of 3.3. The clarity of the chatbot's error explanations received a slightly higher average score of 3.4. Notably, most participants agreed that the chatbot helped improve certain aspects of their English, with five out of seven giving a score of 4 for this question. Additionally, when asked whether they would be interested in using a similar chatbot in future video conferences, all participants but one gave scores of 4 or 5, demonstrating a general interest in this kind of tools.

The audio recordings from the online meetings, descriptions of the role-playing activities, and the corresponding transcriptions are available as part of the English Learners Role-Playing Dialogue Dataset (ELRD), released under a CC license.[27]

Although we do not plan to involve human English teachers in the near future to evaluate the system's error detection capabilities or the interaction between chatbot and students from a teacher's perspective, we are considering this for later stages.

5 Conclusions

In this paper, DeMINT, an innovative conversational intelligent tutoring system designed to enhance English proficiency of non-native speakers through the analysis of online meeting transcriptions, has been presented. Our system leverages the latest advancements in LLMs and integrates various techniques such as in-context learning, retrieval augmented generation, grammatical error correction, and error-preserving speech synthesis and generation. We have provided a comprehensive overview of the system's architecture, including modules for diarization, speech recognition, error correction and annotation, error explanation, knowledge tracing and chatbot orchestration. A pilot evaluation of the system's effectiveness through controlled interactions with L2-English students has been carried out utilizing role-playing games to simulate real-life conversations. Our ultimate goal is to create a scalable, accessible tool that mimics the guidance of a human tutor, providing personalized and context-aware feedback to help non-native speakers improve their language skills by conveniently leveraging their everyday interactions in English. The code, data, and models developed for this project have been openly released across various repositories to promote further research in the field. The central code repository[28] contains links to the additional datasets and models.

Despite being a work-in-progress, we already foresee some future developments. Potential enhancements include supporting voice cloning with tools such as XTTS-v2 (Casanova et al., 2024) so that the error-preserving STT model can be fine-tuned with each user's voice before the debriefing session. Another line of future research involves integrating conversational interaction with users through speech, thus helping students to improve not only their grammatical skills but also their pronunciation. Most components of the system will likely benefit from new emerging models and techniques; for example, for the error explanation module, very recent end-to-end systems that provide error explanations such as xTower (Treviso et al., 2024) are worth exploring. Additionally, multimodal models could be investigated to integrate the non-verbal aspects of online meetings, such as facial expressions and body language. Finally, another area of future work is to conduct an ablation study to determine the relevance of each component within the overall system and explore their potential replacement by a more advanced prompting strategy on the final LLM model.

Acknowledgments. DeMINT (Automated Language Debriefing for English Learners via AI Chatbot Analysis of Meeting Transcripts) is a project funded via FSTP (financial support to third parties), a mechanism by the European Union to support smaller projects through grants provided by larger, EU-funded initiatives. DEMINT is funded under the UTTER[29] (Unified Transcription and Translation for Extended Reality) project, a collaborative Research and Innovation project under Horizon Europe, grant agreement 101070631.

[27] https://github.com/transducens/elrd

[28] https://github.com/transducens/demint

[29] https://he-utter.eu/

Ethics

Since the human evaluation involves collecting and distributing data from participants, special care has been taken to adhere to relevant ethical guidelines[30] and applicable data protection laws. Specifically, the research ethics committee of our university has overseen the experimental process. Each participant was informed about how their interaction with the model would be used and disseminated, and they signed a consent form. Additionally, participants' personal information has been pseudonymized in the released data.

Limitations

Our current system has several limitations that we are aware of. First, the system is currently designed to work with L1-Spanish/L2-English students. Although the system could be adapted for other languages, this would require additional fine-tuning of the models and the incorporation of language-specific resources. Additionally, the system is currently designed to provide feedback on grammar errors and language usage, but it does not address other aspects of language learning such as vocabulary acquisition or pronunciation. Achieving fluency in the communication with the chatbot poses a significant challenge, and the system may fall short of reaching the spontaneity of a human tutor. Finally, some users may prefer reviewing a report over interacting with a chatbot, as non-native speakers are often aware of many errors caused by the improvisation required during conversation, which they would not make in writing.

A Fine-tuning hyperparameters

Empathetic teacher. To fine-tune Llama-3.1-8B to function as the generic teacher described in Section 3.7, we employed the parameter-efficient 8-bit QLoRA method (Dettmers et al., 2023) using a single A100 GPU with 80 GB of VRAM and the LLaMA-Factory toolkit.[31] The LoRA configuration was set to $r = 8$, $\alpha = 16$, with no dropout applied, and targeting all linear modules. Flash Attention version 2 was used (Dao, 2023), and the sequence length was limited to 4 096 tokens. The learning rate was set to 10^{-4} and then adjusted with a linear learning rate scheduler with 10 warmup steps. The training batch size was 12, and weigths were updated after each minibatch. We used the AdamW optimizer with hyperparameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$, while capping the maximum gradient norm at 1.0. The best model was obtained after 2 900 training steps, achieving a cross-entropy loss of 1.83.

	FT	D-FT	W	D-W
Our test set	31.47	38.81	41.82	39.48
Peoples Speech	47.05	30.77	39.45	40.02
Parler-tts	13.70	15.93	26.26	8.63
mls-eng-10k	13.89	15.37	7.34	8.11
Fleurs	13.12	14.98	16.83	17.43

Table 2: WER results on test sets for the best fine-tuned models and original Whisper models. FT: fine-tuned Whisper, D-FT: fine-tuned distilled Whisper, W: original Whisper, D-W: distilled Whisper.

Error-preserving speech-to-text model. As regards the error-preserving speech-to-text model discussed in Section 3.2, we employed a fine-tuning approach using LoRA (Hu et al., 2022) and some specific training arguments to fine-tune the original Whisper model[32] and one distilled version.[33] The configuration for LoRA was set with $r = 16$, $\alpha = 32$, targeting the modules `q_proj` and `v_proj`. Additionally, no dropout was applied, and no bias was included. We fine-tuned the models on one GPU RTX A6000 with 48 GB of VRAM. For the training arguments, the training batch size was set to 8 for the original model and 28 for the distilled one (its smaller size allowed for a larger batch size). Parameters were updated after each minibatch. We used the Adam optimizer with hyperparameters $\beta_1 = 0.9, \beta_2 = 0.999$ and $\epsilon = 1e - 8$, and a linear learning rate scheduler with 50 warmup steps. The learning rate was set to 10^{-5}. The fine-tuning was run for 7 500 steps in the case the original Whisper model, and 7 000 steps in the case of the distilled one. The model parameters were saved every 500 steps, and evaluations were also conducted every 500 steps. At the end of the training, the best model was chosen based on the lowest word error rate (WER) upon the validation set.[34] Table 2 shows the scores of the best models on different test sets.

[30] https://www.acm.org/code-of-ethics

[31] https://github.com/hiyouga/LLaMA-Factory

[32] https://huggingface.co/openai/whisper-large-v3

[33] https://huggingface.co/distil-whisper/distil-large-v3

[34] The selected models had a WER of 12.14 for the original Whisper model and 18.10 for the distilled one.

References

Nathalie Aichhorn and Jonas Puck. 2017. "I just don't feel comfortable speaking English": Foreign language anxiety as a catalyst for spoken-language barriers in MNCs. *International Business Review*, 26(4):749–763.

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

Hervé Bredin. 2023. pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *Proc. INTERSPEECH 2023*.

Christopher Bryant, Mariano Felice, Øistein E. Andersen, and Ted Briscoe. 2019. The BEA-2019 shared task on grammatical error correction. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–75, Florence, Italy. Association for Computational Linguistics.

Christopher Bryant, Mariano Felice, and Ted Briscoe. 2017. Automatic annotation and evaluation of error types for grammatical error correction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 793–805, Vancouver, Canada. Association for Computational Linguistics.

Christopher Bryant, Zheng Yuan, Muhammad Reza Qorib, Hannan Cao, Hwee Tou Ng, and Ted Briscoe. 2023. Grammatical Error Correction: A Survey of the State of the Art. *Computational Linguistics*, 49(3):643–701.

Andrew Caines, Helen Yannakoudakis, Helen Allen, Pascual Pérez-Paredes, Bill Byrne, and Paula Buttery. 2022. The Teacher-Student Chatroom Corpus version 2: more lessons, new annotation, automatic detection of sequence shifts. In *Proceedings of the 11th Workshop on NLP for Computer Assisted Language Learning*, pages 23–35, Louvain-la-Neuve, Belgium. LiU Electronic Press.

Edresson Casanova, Kelly Davis, Eren Gölge, Görkem Göknar, Iulian Gulea, Logan Hart, Aya Aljafari, Joshua Meyer, Reuben Morais, Samuel Olayemi, and Julian Weber. 2024. XTTS: a massively multilingual zero-shot text-to-speech model. arXiv:2406.04904 [eess.AS].

Anna Izabela Cislowska and Beatriz Peña Acuña. 2024. Integration of chatbots in additional language education: A systematic review. *European Journal of Educational Research*, 13(4):1607–1625.

Daniel Dahlmeier, Hwee Tou Ng, and Siew Mei Wu. 2013. Building a large annotated corpus of learner English: The NUS corpus of learner English. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 22–31, Atlanta, Georgia. Association for Computational Linguistics.

Tri Dao. 2023. Flashattention-2: Faster attention with better parallelism and work partitioning. arXiv:2307.08691 [cs.LG].

Dorottya Demszky, Jing Liu, Zid Mancenido, Julie Cohen, Heather Hill, Dan Jurafsky, and Tatsunori Hashimoto. 2021. Measuring Conversational Uptake: A Case Study on Student-Teacher Interactions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115. Curran Associates, Inc.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The Faiss library. arXiv:2401.08281 [cs.LG].

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.

Yuejiao Fei, Leyang Cui, Sen Yang, Wai Lam, Zhenzhong Lan, and Shuming Shi. 2023. Enhancing grammatical error correction systems with explanations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7489–7501, Toronto, Canada. Association for Computational Linguistics.

Mariano Felice, Christopher Bryant, and Ted Briscoe. 2016. Automatic extraction of learner errors in ESL sentences using linguistically enhanced alignments. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 825–835, Osaka, Japan. The COLING 2016 Organizing Committee.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Weijiao Huang, Khe Foon Hew, and Luke K. Fryer. 2022. Chatbots for language learning: Are they really useful? a systematic review of chatbot-supported language learning. *Journal of Computer Assisted Learning*, 38(1):237–257.

Jaeho Jeon. 2024. Exploring AI chatbot affordances in the EFL classroom: young learners' experiences and perspectives. *Computer Assisted Language Learning*, 37(1-2):1–26.

Jiyou Jia. 2009. CSIEC: A computer assisted English learning chatbot based on textual knowledge and reasoning. *Knowledge-Based Systems*, 22(4):249–255.

Masahiro Kaneko and Naoaki Okazaki. 2024. Controlled generation with prompt insertion for natural language explanations in grammatical error correction. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3955–3961, Torino, Italia. ELRA and ICCL.

Omar Khattab, Keshav Santhanam, Xiang Lisa Li, David Hall, Percy Liang, Christopher Potts, and Matei Zaharia. 2022. Demonstrate-Search-Predict: Composing retrieval and language models for knowledge-intensive NLP. arXiv:2212.14024 [cs.CL].

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:2310.03714 [cs.CL].

Lasha Labadze, Maya Grigolia, and Lela Machaidze. 2024. Role of AI chatbots in education: systematic literature review. *International Journal of Educational Technology in Higher Education*, 21(1):28.

Vladimir I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10:707–710.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Yinghao Aaron Li, Cong Han, Vinay Raghavan, Gavin Mischler, and Nima Mesgarani. 2023. StyleTTS 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 19594–19621. Curran Associates, Inc.

Cristóbal Lozano, Ana Díaz-Negrillo, and Marcus Callies. 2020. Designing and compiling a learner corpus of written and spoken narratives: COREFL. In Christiane Bongartz and Jacopo Torregrossa, editors, *What's in a Narrative? Variation in Story-Telling at the Interface between Language and Literacy*, pages 21–46. Peter Lang.

Jakub Macina, Nico Daheim, Sankalan Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2023. MathDial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5602–5621, Singapore. Association for Computational Linguistics.

Janick Michot, Manuela Hürlimann, Jan Deriu, Luzia Sauer, Katsiaryna Mlynchyk, and Mark Cieliebak. 2024. Error-preserving Automatic Speech Recognition of Young English Learners' Language. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6444–6454, Bangkok, Thailand. Association for Computational Linguistics.

Marvin Minsky. 1986. *The Society of Mind*. Simon & Schuster, New York.

Tomoya Mizumoto, Mamoru Komachi, Masaaki Nagata, and Yuji Matsumoto. 2011. Mining revision log of language learning sns for automated Japanese error correction of second language learners. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP)*, pages 147–155.

Amr M. Mohamed. 2024. Exploring the potential of an AI-based chatbot (ChatGPT) in enhancing English as a foreign language (EFL) teaching: perceptions of EFL faculty members. *Education and Information Technologies*, 29(3):3195–3217.

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault. 2017. JFLEG: A fluency corpus and benchmark for grammatical error correction. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 229–234, Valencia, Spain. Association for Computational Linguistics.

Catherine Nickerson. 2005. English as a lingua franca in international business contexts. *English for Specific Purposes*, 24(4):367–380.

Kostiantyn Omelianchuk, Andrii Liubonko, Oleksandr Skurzhanskyi, Artem Chernodub, Oleksandr Korniienko, and Igor Samokhin. 2024. Pillars of grammatical error correction: Comprehensive inspection of contemporary approaches in the era of large language models. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 17–33, Mexico City, Mexico. Association for Computational Linguistics.

Larry Pitts. 2015. *ESL Role Plays: 50 Engaging Role Plays for ESL and EFL Classes*. ECQ Publishing.

Alexis Plaquet and Hervé Bredin. 2023. Powerset multi-class cross entropy loss for neural speaker diarization. In *Proc. INTERSPEECH 2023*.

Muhammad Reza Qorib and Hwee Tou Ng. 2023. System combination via quality estimation for grammatical error correction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12746–12759, Singapore. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Chrysi Rapanta, Cláudia Gonçalves, João Rui Pereira, Dilar Cascalheira, Beatriz Gil, Rita Morais, Anna Čermáková, Julia Peck, Benjamin Brummernhenrich, Regina Jucks, Mercè Garcia-Milà, Andrea Miralda-Banda, José Luna, Maria Vrikki, Maria Evagorou, and Fabrizio Macagno. 2021. Multicultural classroom discourse dataset on teachers' and students' dialogic empathy. *Data in Brief*, 39:107518.

Keshav Santhanam, Omar Khattab, Christopher Potts, and Matei Zaharia. 2022a. PLAID: An efficient engine for late interaction retrieval. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, page 1747–1756, New York, NY, USA. Association for Computing Machinery.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022b. ColBERTv2: Effective and efficient retrieval via lightweight late interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3715–3734, Seattle, United States. Association for Computational Linguistics.

Maganat Shegebayev. 2023. Rise of English as business lingua franca at the turn of the century: An overview. In Stanley D. Brunn and Roland Kehrein, editors, *Language, Society and the State in a Changing World*, pages 357–365. Springer International Publishing, Cham.

Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, Morteza Behrooz, William Ngan, Spencer Poff, Naman Goyal, Arthur Szlam, Y-Lan Boureau, Melanie Kambadur, and Jason Weston. 2022. BlenderBot 3: a deployed conversational agent that continually learns to responsibly engage.

Yixiao Song, Kalpesh Krishna, Rajesh Bhatt, Kevin Gimpel, and Mohit Iyyer. 2024. GEE! grammar error explanation with large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 754–781, Mexico City, Mexico. Association for Computational Linguistics.

Felix Stahlberg and Shankar Kumar. 2021. Synthetic data generation for grammatical error correction with tagged corruption models. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 37–47, Online. Association for Computational Linguistics.

Katherine Stasaski, Kimberly Kao, and Marti A. Hearst. 2020. CIMA: A large open access dialogue dataset for tutoring. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 52–64, Seattle, WA, USA → Online. Association for Computational Linguistics.

Toshikazu Tajiri, Mamoru Komachi, and Yuji Matsumoto. 2012. Tense and aspect error correction for esl learners using global context. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 198–202.

Yuki Takashima, Yusuke Fujita, Shinji Watanabe, Shota Horiguchi, Paola García, and Kenji Nagamatsu. 2021. End-to-end speaker diarization conditioned on speech activity and overlap detection. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 849–856.

Marcos Treviso, Nuno M. Guerreiro, Sweta Agrawal, Ricardo Rei, José Pombal, Tania Vaz, Helena Wu, Beatriz Silva, Daan van Stigt, and André F. T. Martins. 2024. xTower: A multilingual LLM for explaining and correcting translation errors. `arXiv:2406.19482 [cs.CL]`.

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.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 180–189, Portland, Oregon, USA. Association for Computational Linguistics.

Shunan Zhang, Cheng Shan, John Sie Yuen Lee, ShaoPeng Che, and Jang Hyun Kim. 2023. Effect of chatbot-assisted language learning: A meta-analysis. *Education and Information Technologies*, 28(11):15223–15243.

Mingchen Zhuge, Haozhe Liu, Francesco Faccio, Dylan R. Ashley, Róbert Csordás, Anand Gopalakrishnan, Abdullah Hamdi, Hasan Abed Al Kader

Hammoud, Vincent Herrmann, Kazuki Irie, Louis Kirsch, Bing Li, Guohao Li, Shuming Liu, Jinjie Mai, Piotr Piękos, Aditya Ramesh, Imanol Schlag, Weimin Shi, Aleksandar Stanić, Wenyi Wang, Yuhui Wang, Mengmeng Xu, Deng-Ping Fan, Bernard Ghanem, and Jürgen Schmidhuber. 2023. Mindstorms in natural language-based societies of mind. In *R0-FoMo: Robustness of Few-shot and Zero-shot Learning in Foundation Models, NeurIPS 2023*.