

ISSUES

POTENTIAL AFFORDANCES of GENERATIVE AI in LANGUAGE EDUCATION: DEMONSTRATIONS and an EVALUATIVE FRAMEWORK

Austin Pack¹^a, Jeffrey Maloney¹^b

¹ Faculty of Education and Social Work, Brigham Young University-Hawaii

Keywords: Artificial Intelligence, Language Education, Materials Creation, Language Assessment

<https://doi.org/10.56297/BUKA4060/VRRO1747>

Teaching English with Technology

Vol. 23, Issue 2, 2023

With recent public access to large language models via chatbots, the field of language education is seeing unprecedented levels of interest in how AI will affect language learning and teaching. As attention is primarily focused on student misuse of the technology, the potential affordances of generative AI tools may often be overlooked. In this paper we highlight potential affordances of AI generative tools, such as ChatGPT, for creating language learning and teaching materials and assessments. We offer suggestions on crafting prompts as well as provide example prompts and the materials and assessments generated by ChatGPT in response. We also present a framework for evaluation that combines both of Hubbard's (1988) framework for CALL evaluation and The Douglas Fir Group's (2016) transdisciplinary framework that may assist language educators in determining when it is appropriate to leverage generative AI tools. We then briefly illustrate how the framework has been used in our own evaluations. Finally, we suggest language educators consider how the concerns and opinions of using AI-based tools held by individuals and institutions within their own micro, meso, and macro sociocultural levels align or conflict with those of the individuals and institutions of their students' micro, meso, and macro sociocultural levels.

1. Introduction

While artificial intelligence (AI) has been used in the field of language education in various forms for some time, recent public open access to large language models (LLMs) via AI chatbots, such as OpenAI's ChatGPT, Google's Bard, and Microsoft's Bing, has resulted in unprecedented levels of interest in how AI may disrupt education in general and language learning and teaching more specifically. In recent years AI has been used in a variety of ways in language education, including automated written corrective feedback (AWCF; Koltovskaia, 2020), machine translation (Paterson, 2022), conversation practice (Lorentzen & Bonner, 2023), generating lesson ideas, presentation notes, and texts for reading (Bonner et al., 2023), and automatic text generation (Godwin-Jones, 2022). LLMs however, have drawn special attention in recent years due to their capabilities. Open AI's GPT-3 (third-generation Generative Pre-trained Transformer) for example, has the ability to answer questions, challenge faulty premises, admit mistakes, and produce unique writing better than what many individuals can produce (Elkins & Chun, 2020; Floridi & Chiriatti, 2020). AI Chatbots serve as a way for users to

^a austin.pack@byuh.edu

^b jeffrey.maloney@byuh.edu

interface with the LLMs to generate text and complete tasks (hence the reason why these technologies are often referred to as generative AI). ChatGPT, for example, has been used to pass a Wharton MBA exam (Mollman, 2023) and publish academic journal articles (e.g., King, 2023). As such, universities, academics, and educators have expressed alarm concerning student use of generative AI and are wrestling with how to adjust teaching, assessment, and policies (Ferlazzo, 2023; Huang, 2023) to prevent students from abusing the technology by having AI generate unique texts that are far above the students' abilities. Most of the ongoing discussions related to the use of generative AI in education have focused on the possible ways students may (ab)use the technology and related ethical considerations. While such concerns and related discussions are valid and merited, we suggest that the potential affordances generative AI may offer language educators are often overlooked, and more attention needs to be given to the work of educators and researchers who explore the positive potentials of these technologies (e.g., Bonner et al., 2023).

The purpose of this paper, therefore, is to draw attention to additional potential uses of AI text generating tools, such as chatbots and LLMs, for creating language learning and teaching materials and assessment and to offer a framework to assist language educators in making informed decisions as to if, when, and how to use these tools. As there are a growing number of LLMs and chatbots, and these undergo regular updates, it is beyond the scope of this paper to offer up a comparison of the chatbots and their corresponding LLMs. Instead, we have opted to focus on ChatGPT (GPT-3.5), as it is likely the most widely known and used generative AI tool. We begin by offering a few suggestions on crafting prompts for generative AI. We then demonstrate how ChatGPT (GPT-3.5) can be used to generate discussion questions, modify material to be appropriate for English language learners of different proficiency levels, and generate other text such as models of writing, handouts with explanations and/or practice exercises, transcripts for listening assessments, and creating and using rubrics for evaluating students' writing. While ChatGPT (GPT-3.5) can be used in additional ways in creating text for materials and assessments, due to limited space we are unable to make an exhaustive list. Instead, we encourage language educators to use the examples presented here as a springboard to create and revise AI generated texts of their own that may be of use in their particular teaching contexts. After providing several demonstrations we then present a potential framework for instructors to consider adopting, which combines elements of Hubbard's (1988, 2021) framework for evaluation and Bronfenbrenner's (1979) framework on the ecology of human development. By combining the elements of teacher and learner fit with considerations across micro, meso, and macro-levels, instructors can use the extended framework to assist in confronting the potential points of concern and conflict that may exist.

2. Getting started

ChatGPT(GPT-3.5) generates text based on prompts provided by users. In general, we have found the output to be more useful in creating language learning and teaching materials and assessments when the user is specific in their request. Consider the difference between a basic prompt and one that we tailored specific to our students:

Example general prompt:

Create a handout for English language learners that helps them learn and practice using passive voice.

Example specific prompt:

You are a professional English language teacher. Create a handout for CEFR B1 English language learners that helps them learn and practice using passive voice. It should begin with a simple explanation of what passive voice is and how it is used in academic writing. Provide a couple of model sentences. Then provide 3 series of practice exercises. Each series should contain 5 exercises and each series should be scaffolded, progressing from easy to difficult. Use the following words throughout the handout: employ, transition, isolated, insist on, preservation, accustomed to, assemble, contradiction, principle, pursue.

Just as with any kind of educational technology, the more familiar and practiced one is using generative AI tools, the more likely he/she will find them empowering for use in the classroom. We encourage practitioners to experiment with the technology and to explore its limitations and potentials. While the example prompts provided throughout this article may serve as an initial guide, writing prompts that yield desired results is a skill that can be developed through continued use of the technology. We therefore encourage readers to try including different levels of specificity in their prompts by including (or not including) information regarding the topic of interest (e.g., tradition and progress), learner proficiency level (e.g., CEFR A1-C2), vocabulary words (e.g., from a particular unit or chapter), grammar patterns the practitioner wants students to notice (e.g., adjective clauses), the style or voice of writing (e.g., academic), the length of the text, and the number of and difficulty of questions and/or exercises.

We suggest four simple steps readers can follow to develop effective prompts. First, assign a role or identity to the chatbot (“You are a professional English writing teacher”). Next, define an objective or purpose for the chatbot (“You will hold a writing conference with me and help me improve my writing”). Then define the bounds within which you want the chatbot to respond by specifying the context and constraints. More detail is usually better (“I am writing a paper on sustainability and tourism in Hawaii. Don’t fix my writing

for me. Ask me guiding questions to help me fix my own writing.”). Lastly, and perhaps most importantly, refine the output through continued conversation (“Consolidate into a table all the grammar mistakes in my writing that we have discussed today. In the first column, write the mistake, in the second column, identify what kind of grammar mistake was made, in the third column, explain why what I had written is wrong. Now based on the mistakes I have made, make a grammar handout that addresses my grammar mistakes with practice exercises”)

Worth noting, if educators want language learners to interact directly with AI chatbots, Poe ([Poe.com](https://poe.com)) allows users to create their own tailored chatbots (based on GPT-3.5, GPT-4, or Claude) and stipulate how they want these chatbots to function for other users. These can be shared via the internet. In other words, a language teacher could use this platform to assign a role, purpose, and context/constraints to a chatbot, and then share that tailored chatbot with their language learners, so that the students do not need to assign a role, purpose and context to the chatbot.

3. Creating materials using ChatGPT(GPT-3.5)

Below we provide example prompts for creating materials and ChatGPT’s responses. Due to constraints in space, we share only some of ChatGPT’s responses.

3.1. Generating discussion questions

Our prompt:

Write four one-minute timed speaking response prompts for CEFR B2 English language learners on the topic of tradition and progress. Use some of the following vocabulary in the prompts: employ, transition, isolated, insist on, preservation, accustomed to, assemble, contradiction, principle, pursue

ChatGPT’s response:

- 1.“Discuss the importance of preservation of traditional customs and practices in today’s society.”
- 2.“How do you think the transition from traditional to modern ways of life affects individuals and communities?”
- 3.“In your opinion, is it possible to balance the pursuit of progress with the preservation of tradition?”
- 4.“How do you think isolated communities maintain their traditional customs and practices in the face of modern influences?”

Commentary:

We have found ChatGPT as a useful resource in quickly generating ideas for discussion questions, but without further refinement the questions generated tend to be overly broad. It may be useful to enhance the questions by having ChatGPT revise discussion questions to include specific vocabulary, or by refining them with additional prompts that provide additional context (e.g., “Now revise the questions so that they relate to the lives of college students in Hawaii” or “Now revise the questions so that they are about tradition and progress in Japan, where the language learners are from”).

3.2. Modifying material to be appropriate for learners of different proficiency levels

Bonner et al. (2023) have pointed out how ChatGPT can be leveraged to summarize text in level appropriate language that is more digestible. We suggest that ChatGPT’s ability to generate text according to different proficiency levels can be used more broadly, such as in creating discussion questions of different levels, and in providing scaffolded models of writing, that enable language educators to support students of varying proficiency levels who are in the same class.

Our prompt:

Take the following question and write it in several ways. Write it once for each CEFR level¹ (CEFR A1, A2, B1, B2, C1 and C2 levels): “What is the importance of keeping traditional customs and practices in today’s society?”

ChatGPT’s response:

CEFR A1: “What is tradition?”

CEFR A2: “Why is tradition important?”

CEFR B1: “What is the importance of keeping traditional customs and practices in today’s society?”

CEFR B2: “What role do you believe traditional customs and practices should play in contemporary society, and how can they be effectively preserved?”

CEFR C1: “What are the potential implications of neglecting traditional customs and practices for contemporary society and how can they be effectively preserved and integrated into modern life?”

CEFR C2: “What are the potential implications of neglecting traditional customs and practices for contemporary society in

¹ CEFR stands for Common European Framework of Reference for Languages. A1 & A2 stand for basic user, B1 & B2 stand for independent user, and C1 & C2 stand for proficient user, with 1 being low and 2 being high.

terms of cultural heritage, identity and social cohesion, and what strategies could be employed to effectively preserve and integrate them into modern life?”

Commentary:

It may be prudent, especially for longer texts, for educators to verify that ChatGPT is indeed generating text at the desired proficiency level. This can be done by means of a text analyzer like Cathoven (<https://hub.cathoven.com/>) or [EnglishGrammar.Pro](https://englishgrammar.pro/test.php) (<https://englishgrammar.pro/test.php>). Generally speaking, we have found GPT-4 to perform better than GPT-3.5 in adhering to vocabulary of a particular proficiency level.

3.3. Generating model writing samples for comparison

Our initial prompt:

Write two paragraphs (approximately 200 words in length) for CEFR B1 English language learners on the topic of progress and environmental degradation. The paragraphs should have the same content, but be different in voice; the first paragraph should use common everyday English, and the second paragraph should use academic English.

Our follow-up prompt:

Create four exercise questions that help students to notice the difference in voice between the two paragraphs.

Commentary:

Using ChatGPT in this way is beneficial because it allows students to compare different styles and/or proficiency levels of writing. This allows educators to generate tailored text that is slightly beyond students' current level of competence and can help offer learners with opportunities to actively notice key differences in features of language between models, which the extant literature suggests to be of great import (e.g., Krashen, 1982; Schmidt, 1990).

3.4 Generating handouts with explanations or practice exercises

Our initial prompt:

Create a handout for CEFR B2 English language learners on the use of academic transition words and phrases. The handout should include an explanation and a list of academic transition words and phrases (organized according to function).

ChatGPT's response (formatted for easier readability):

Academic Transition Words and Phrases Handout for CEFR B2 English Language Learners

Introduction:

Academic writing requires a clear and logical flow of ideas. One way to achieve this is by using transition words and phrases to connect and link different parts of your text. These words and phrases help to indicate the relationship between different ideas and provide a sense of coherence to your writing.

List of Academic Transition Words and Phrases:

To indicate addition:

Furthermore

Additionally

Moreover

Besides

Also

And

In addition

As well as

Afterwards

Recently

[ChatGPT then listed eight transition words for each of the following categories: indicating contrast, indicating time, cause and effect, similarity and difference, examples, generalization, and emphasis.]

Our follow-up prompt:

Provide a practice exercise for each category of function.

Commentary:

ChatGPT can be used to generate different types of handouts, such as worksheets and study guides. A useful feature is consolidating information into tables. Users can describe what should belong in each column and/or row and ChatGPT will automatically generate a table according to the specifications desired.

4. Creating assessments using ChatGPT(GPT-3.5)

Below we provide example prompts for language assessment and ChatGPT's responses. Due to constraints in space, we share only some of ChatGPT's responses.

4.1 Generating transcripts for listening assessments

Bonner et al. (2023) have demonstrated how ChatGPT can be used for generating reading passages and related questions for assessment. Below we expand on this idea and demonstrate how ChatGPT can assist in developing a transcript for a listening assessment.

Our prompt:

You are an expert on language assessment. You will help me create a listening test for CEFR B1 English language learners enrolled in an Academic English class. Create a transcript for a 1000-word podcast interview between the host and a professor of psychology. The topic of the interview should be about what happens to the human mind when one's culture and traditions are being replaced by new or foreign culture and traditions.

ChatGPT's response:

Host: Hello and welcome to the podcast. Today, we have a special guest with us, Professor Jane Smith from the Department of Psychology at XYZ University. Our topic of discussion for today is how the human mind reacts to changes in cultural and traditional values. Welcome to the show, Professor.

Professor: Thank you for having me.

Host: So, Professor, can you please explain to our listeners how the human mind reacts when a person's cultural and traditional values are being replaced by new or foreign values?

Commentary:

After refining the transcript through continued conversation and additional prompts we used ElevenLabs' Speech Synthesis text-to-speech generator (<https://beta.elevenlabs.io/speech-synthesis>) to render AI generated realistic human speech voicing for the host of the podcast and the professor being interviewed, thereby saving us time and effort in finding volunteer voice actors and recording audio. The example prompt and resultant transcript demonstrate how generative AI can be used to link together the content of different thematic units covered in class. For example, in our academic English listening and speaking class the first unit is about tradition and progress, while

the second unit is about the human mind. Using ChatGPT allowed us to generate a transcript that brought these two different themes together in an authentic and meaningful way.

It is important to note that LLMs are known to at times confidently make up information that may not be true. For example, LLMs may fabricate academic references that are convincing. These instances where LLMs generate incorrect information are known as “hallucinations”. If the purpose of the language assessment is solely to measure student’s comprehension of information conveyed, then these hallucinations are likely unimportant. If, however, the goal of the listening is to convey knowledge and information, then hallucinations are highly problematic. In either case, users should use LLMs for assessment with caution, and any text generated should be carefully and thoroughly scrutinized.

4.2. Creating rubrics

Our prompt:

Create a rubric for assessing English language learners’ writing. The rubric should have four categories (content, organization, language use, and use of sources and evidence). Use a scale of 1-10 points for each category, with descriptors for points 1-5, 6-7, 8-9, and 10.

ChatGPT’s response:

Content (10 points possible)

1-5: Minimal understanding of topic, little or no relevant information provided.

6-7: Basic understanding of topic, some relevant information provided.

8-9: Good understanding of topic, clear and detailed information provided.

10: Excellent understanding of topic, clear and detailed information provided with insightful analysis.

Organization (10 points possible)

1-5: Poor organization, difficult to follow the main ideas.

6-7: Basic organization, main ideas are clear but may lack coherence.

8-9: Good organization, clear and logical flow of ideas.

10: Excellent organization, clear and logical flow of ideas with smooth transitions.

Language Use (10 points possible)

1-5: Significant errors in grammar, vocabulary, and/or spelling that impede understanding.

6-7: Some errors in grammar, vocabulary, and/or spelling that may impede understanding.

8-9: Few errors in grammar, vocabulary, and/or spelling that do not impede understanding.

10: No errors in grammar, vocabulary, and/or spelling.

Use of Sources and Evidence (10 points possible)

1-5: No sources or evidence used, or sources used are not relevant to the topic.

6-7: Sources and evidence used but not cited or not used effectively.

8-9: Relevant sources used and cited effectively to support main ideas.

10: Relevant sources used and cited effectively to support main ideas with insightful analysis of the sources.

Commentary:

A variety of rubrics can be made with ChatGPT, including analytic, developmental, holistic, and checklists. In addition to using LLMs to create rubrics, educators can ask for LLMs to critique rubrics and provide suggestions for improvement.

4.6. Evaluating student writing based on a rubric

Our prompt:

Evaluate the following students writing (A) according to the following rubric (B).

A: Culture shock is a feeling of disorient that occur when somone is in a new culture. It can be caused by factors such as differences in language, culture, and ideas and culture shock include homesickness, difficulty eat to new foods, and feeling of isolation or depression and it is important to recognize that culture shock is normal part of adjusting to new culture. And it is possible to overcome it with time and effort. Someway to cope with culture shock is connect with friends and family back home,

and get out support from others who also have same feelings. By choose to learn about and adapt to a new culture, individuals can slowly learn enjoy and appreciate new culture.

B: [We copied and pasted the rubric above that ChatGPT generated.]

ChatGPT's response:

Content (7/10 points)

Organization (8/10 points)

Language Use (6/10 points)

Use of Sources and Evidence (6/10 points)

The student has a good understanding of the topic of culture shock and provides relevant information, but could be more detailed. The organization of the writing is clear and logical, but could be improved with smoother transitions. There are some errors in grammar, vocabulary, and spelling that impede understanding, but the writing is still understandable. The use of sources and evidence is limited and not effectively cited or used to support the main ideas.

Our follow-up prompt:

Provide specific suggestions to the student regarding how they can improve the paragraph. Use specific examples from their paragraph.

Commentary:

While the application of AI in assessing English writing is not new, the use of LLMs and the public's access to this technology is. As the reliability and validity of using generative AI for assessing writing has yet to be empirically tested, we strongly urge educators to use discretion and caution in using generative AI in this way. In our own exploration of using generative AI for language assessment, we have found GPT-4 to be more reliable than GPT-3.5 and Bard. With careful prompt engineering we have found GPT-4 to be as equally reliable as human raters within our institution. We have found that generative AI tools tend to perform better when using holistic rubrics that are simpler than analytic rubrics. We have also found success in using GPT-4 to convert an analytic rubric to a holistic rubric and then apply the holistic rubric to students' writing.

5. Discussion and conclusion of example uses

The above examples are just some potential ways ChatGPT could assist language practitioners; there are additional ways of using ChatGPT that we cannot demonstrate due to restrictions in space, such as creating writing prompts and applying universal design for learning to materials, amongst others. The materials and assessments generated by ChatGPT are not without their flaws. Language educators should not consider ChatGPT or any other AI-based tools to function as an autopilot that requires no oversight. While generative AI can undertake a substantial portion of the preliminary tasks in materials and assessment creation, the text these tools produce is often not fit for educational purposes until it has been scrutinized, revised, and refined by a user. Perhaps generative AI tools will be most useful when teachers lack access to better materials, have insufficient time to prepare materials, or use the generated text as an initial springboard for ideas for material and assessment creation. While it is beyond the scope of this paper to explain in detail all of the issues arising from using AI-based tools in the language classroom, language educators ought to be aware of three common reasons why AI tools raise concerns; these include:

1. AI-generated work may be viewed as unoriginal and therefore equivalent to plagiarism. The reason for this is that the text generated is based on a data set of pre-existing texts from other sources that the AI model was trained on, which are not made known to the user.
2. AI models are prone to biases, as they are trained on limited data sources and may replicate errors present in those sources. AI bias and discrimination issues have been raised in various fields, including employment (Bogen, 2019), healthcare (Obermeyer et al., 2019), criminal justice (Idder & Coulaux, 2021), and education (Baker & Hawn, 2021). The use of AI in education has been recognized as posing a risk of exacerbating various forms of injustice, such as systemic bias and/or unequal access (Akgun & Greenhow, 2022).
3. AI tools' ability to make judgements with limited information and to evaluate and reflect on criteria (e.g., assessment items) may be well below that of an experienced language educator.

6. Framework for utilizing AI in language teaching across contexts

Questions as to if, when, and how to use AI in teaching ought to be carefully considered by language educators. There are multiple frameworks and suggestions that may assist language educators in such decisions relating to the use of educational technology. For example, Chun et al. (2016) provide three areas for consideration, including affordances, student experiences and expectations, and language learning environment. Hubbard's (1998; 2011) framework includes consideration of a tool or site's operational description across both teacher and learner fit. While these frameworks are for software and

websites more generally, there are also additional frameworks for more specific kinds of applications, such as mobile language-learning apps (see Rosell-Aguilar, 2017). There are additional frameworks that have been proposed (for further information on previous frameworks, see McMurry et al., 2016), however, most contain similarities in their approach; they focus on evaluating the potential of a tool or application to address language learning or teaching concerns.

While these frameworks each have their own advantages and disadvantages, the integration of novel AI technology may require different approaches and considerations. Specifically, it is of great import for educators to consider whether the use of these tools would be a good fit in their specific contexts where attitudes towards using AI generative tools and authorship may vary by student, teacher, and educational institution. For this reason, we offer a new framework for utilizing AI in language teaching across contexts. This novel framework is a combination of Hubbard's (1988, 2021) integrated framework for CALL courseware evaluation and The Douglas Fir Group's (2016) transdisciplinary framework on the multifaceted nature of language learning and teaching.

Simply put, Hubbard's (1988, 2021) evaluation framework has five central components: operational description (procedure), learner fit (design), teacher fit (approach), implementation schemes, and appropriateness judgements. While Hubbard (2011) has extended the model to account for the web and other resources, we believe the model can be further adapted to consider the diversity of sociocultural factors that exist across various levels.

The Douglas Fir Group (2016) have presented a transdisciplinary framework, based on Bronfenbrenner's ecological framework for human development (U. Bronfenbrenner, 1979; Urie Bronfenbrenner & Morris, 2007) that highlights the multifaceted nature of language learning and teaching. This includes three levels of mutually dependent influence: the macro level of ideological structures (e.g., belief systems and cultural values, amongst others), the meso level of sociocultural institutions and communities (e.g., schools, social organizations, families), and the micro level of social activity (e.g. individuals interacting with others). While not designed specifically for the evaluation of technology, the transdisciplinary nature of the framework allows for it to be applied when considering the appropriacy of utilizing technology within and across contexts.

We propose that a framework that combines both Hubbard's (1988) framework for CALL evaluation and The Douglas Fir Group's (2016) transdisciplinary framework may assist language educators in determining when it is appropriate to leverage generative AI tools such as LLMs and chatbots (see [Figure 1](#)). The appropriacy of using AI in language education depends largely on the alignment of how well it fits the teacher and learner at the micro, meso, and macro sociocultural levels. Micro-level considerations

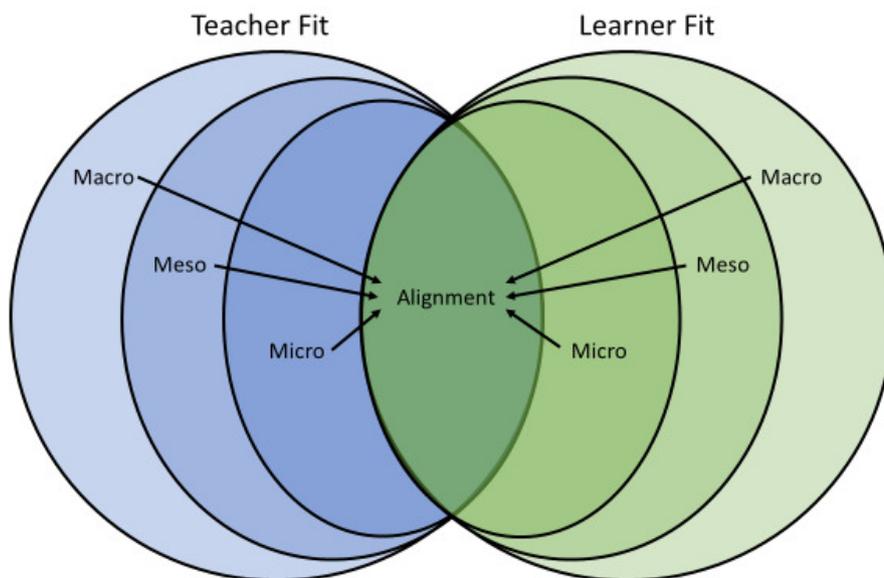


Figure 1. Framework for utilizing generative AI in language teaching across contexts

include interactions amongst individuals and small groups (e.g., teacher and students in a language class). When considering the use of AI at this level, an instructor would be evaluating the tool's appropriacy within the culture of the classroom and even within the individual interactions that an instructor may (or may not have) with students. An instructor would consider whether the use of the tool aligns with the objectives of the course, the abilities, access, and needs that students might have to technology and linguistic resources.

Meso-level considerations include intermediate level organizations (e.g., communities and local institutions). This would include considerations of institutional policy (e.g., academic integrity), access, culture, and overall perspectives that exist across the institutions and communities that are immediately surrounding the classroom. For example, an instructor ought to consider whether there are any official policies regarding use of these tools at the institutional level. Furthermore, evaluating whether the tool is viewed as acceptable or valuable for use by other instructors or administrators can be an important step in deeming whether it would be good to adopt. For learners, consideration of student culture, access to resources, and expectations generally across the campus would be of importance.

The macro level refers to large-scale systems and structures. This would include consideration of the political, economic, and cultural perspectives, and access to the technology that might exist within the larger regional, national or even global scales. For example, consideration of national movements and perspectives within the broader culture may have some import into the

evaluation process. This level would include how teachers are viewed, the state of technological adoption and even ontological and epistemological perspectives within different fields of academia.

Individuals and institutions throughout the micro, meso, and macro levels of society are likely to have differing concerns and opinions about the use of AI generative tools in education. This is especially true for language learning contexts, where teachers, students, and administrators frequently come from different cultural, educational, political, and ideological backgrounds. We suggest that language educators ought to consider how the concerns and opinions of using AI-based tools held by individuals and institutions within their own micro, meso, and macro sociocultural levels align or conflict with those of the individuals and institutions of their students' micro, meso, and macro sociocultural levels.

Evaluation of generative AI tools using this framework can be done step-by-step, beginning at the macro-level, then considering meso-level, and finally micro-level contextual considerations. For example, in judging the appropriacy of using AI in their language class, be it by teacher or students, teachers might ask themselves the following questions, or ones similar:

- Macro-level: Does the proposed use of AI generative tools align with the society-wide ideologies and systems of beliefs of me and my students (e.g., ideas regarding authorship and plagiarism)?
- Macro-level: Do institutions, such as the national and local governments of the country I am teaching in prevent access to web-based AI generative tools?
- Meso-level: Does the proposed use of AI generative tools align with the values, policies, and practices of the intuitions that I and my students are a part of and/or governed by (e.g., academic integrity policies)?
- Micro-level: Does the proposed use of AI generative tools align with my teaching approach and familiarity/competency with technology?
- Micro-level: Does it align with the linguistic needs of my students and their familiarity/competency with technology?
- Micro-level: Is the proposed use of generative AI efficacious in its proposed application within my classroom and teaching??

Opinions toward and policies relating to the use of AI in language education may not always align. In some situations, teachers can align the desired use of AI with the desires, expectations, and policies/laws of individuals and institutions within the various sociocultural levels. For example, a language teacher who has limited time to provide corrective feedback on students

writing may want to use an AI-based tool, such as ChatGPT, to generate corrective feedback. The government of the country in which the teaching is occurring, however, may enforce specific laws aimed at protecting the privacy of learners; submitting student work to a third party (i.e. OpenAI, Google) may violate such laws. This teacher could likely align the use of AI with the expectations of the government by including an “opt in” statement for students whereby students agree to having their work submitted to and evaluated by a LLM.

7. An example application of the framework: Utilizing ChatGPT for English language learning and teaching at a small, private, liberal arts university in the Pacific (United States)

What follows below is a vignette of the application of the framework as it has been applied in consideration of ChatGPT’s use within the authors’ own classrooms and university.

Operational Description: ChatGPT is an AI chatbot developed by OpenAI. It can answer questions, challenge faulty premises, admit mistakes, and produce unique writing. To use this resource, students and teachers need access to the internet on a digital device. Teachers can use ChatGPT to generate ideas and/or materials related to English language learning, teaching, and assessment. Currently ChatGPT(GPT-3.5) is free to use, while ChatGPT(GPT-4) is not.

Macro:

- Teacher Fit:
 - There is great concern that this technology will lead to upheaval in higher education in general and in language learning more specifically. When it is appropriate to use these tools is still being worked out. As such, there is a need for teachers to explore the affordances and drawbacks of the technology, to share best practices and policies, and to engage in discussions as to how the introduction of this technology is affecting language learning, teaching, and research.
 - Our country has laws to protect the educational rights and privacy of students. These laws prevent disclosure of personally identifiable information derived from educational records. If we submit student writing to AI-based tools like Chat GPT, then we should ensure that the student is not identifiable in anyway.
- Learner Fit:
 - Many of our students come from cultures with views of authorship and plagiarism that differ from those common in the university we teach at within the United States. Students

may not even recognize that submitting ChatGPT generated writing as their own violates the university's academic integrity policy. Therefore, student use should be minimal, and used only with strong guidance as to when and how it is appropriate to use (e.g., when the use is helping them develop as a writer rather than doing the writing for them).

Meso: Private, liberal arts university with a majority of students coming from Asia and Pacific rim countries.

- Teacher Fit:
 - While there is an increasing interest amongst faculty members in the use of AI-based tools, most discussions center on student use. The university has yet to adopt a policy for teachers or students specific to the use of AI-based tools for generating text. Use of ChatGPT should be aligned with the academic integrity policy currently in place. We should encourage discussions of potential policies specific to the use of AI-based tools in writing at the program and university levels.
- Learner Fit:
 - We as teachers should be clear on what we consider acceptable/unacceptable use of AI-based tools. We should verify that our students understand what is expected of them regarding academic integrity and what institutional policies are in place.

Micro: Both authors are instructors within EAP and TESOL (pre-service teacher) programs, with classes in both areas.

- Teacher Fit:
 - We view the technology as having great potential for use as a tool to aid with many aspects of our day-to-day tasks in teaching including materials creation and assessment. We are comfortable in implementing new technology in our classes.
- Learner Fit:
 - As our ESL students are B1-B2 CEFR level students, without proper guidance they will likely struggle to use ChatGPT in appropriate ways. If we use texts generated by ChatGPT then we should ensure that these texts have been tailored to the B1-B2 levels.

- Our TESOL pre-service teacher students will encounter both student and teacher use of AI-based tools like ChatGPT when they enter the work force. As such, they should be made aware of these AI-based tools and guidance on how to use them appropriately as students and teachers should be given.

Conclusion: In light of these considerations we deem it appropriate to use ChatGPT as a teacher in material and assessment creation and for pre-service TESOL teachers use with proper guidance. For now, we will strongly caution students about the use of ChatGPT until we have created example documents that highlight appropriate use of carefully selected and vetted prompts. It may be beneficial to create a list of prompts that students can use, or even create our own application or tool that allows students to input text and get feedback from ChatGPT using only the prompts we have already tested.

8. Conclusion

Given most general discussions on the use of generative AI tools in education center on student (mis)use of the technology, this paper has attempted to draw attention to the potential uses of LLMs and chatbots for language educators in creating language learning and teaching materials and assessments. We have provided examples of how this technology can be used to for creating materials and assessment, as well as applying rubrics for evaluating students' writing.

The framework we have proposed will hopefully provide a helpful heuristic with which to consider the multiplicity of potential conflicts and concerns that have arisen regarding the use of generative AI. We anticipate that the outcome of considerations of ChatGPT and other tools using this framework might also be useful for informing discussions of policy within departments and institutions.

Moving forward there is much work that needs to be done to ensure AI-based tools are being used appropriately by students and teachers in language education. One avenue that needs further exploration is identifying additional potential uses beyond the examples given in this paper. We therefore encourage practitioners to experiment with the technology and to contribute to the ongoing discussions within the academy. Second, ChatGPT, while perhaps the most widely known AI chatbot, is not the only AI chatbot available. Google Bard and Microsoft's Bing are also now available to the public. It is likely true that each of these will perform differently in creating language learning and teaching materials and assessments. As such, practitioners and researchers ought to experiment and tease out the advantages and disadvantages of each AI chatbot. As these chatbots are developed iteratively, evaluation of the chatbots should be ongoing. Third, research into the reliability and validity of ChatGPT and other chatbots' responses to prompts aimed at creating

materials and assessments is crucial. Lastly, research into students' perspectives of teacher use of AI tools to generate language learning materials and assessment would be insightful.

Conflict of interest statement

The authors have no sources of funding and no conflicts of interest to declare.

Author contribution statement

Author 1 and author 2 contributed to the writing and revision of this manuscript.

REFERENCES

- Akgun, S., & Greenhow, C. (2022). Artificial intelligence in education: Addressing ethical challenges in K-12 settings. *AI and Ethics*, 2(3), 431–440. <https://doi.org/10.1007/s43681-021-00096-7>
- Baker, R. S., & Hawn, A. (2021). *Algorithmic Bias in Education*. <https://doi.org/10.35542/osf.io/pbmvz>
- Bonner, E., Lege, R., & Frazier, E. (2023). Large language model-based artificial intelligence in the language classroom: Practical ideas for teaching. *Teaching English with Technology*, 2023(1), 23–41. <https://doi.org/10.56297/bkam1691/wieo1749>
- Bronfenbrenner, U. (1979). *The ecology of human development: Experiments by nature and design*. Harvard University Press.
- Bronfenbrenner, Urie, & Morris, P. A. (2007). The bioecological model of human development. In W. Damon & R. Lerner (Eds.), *Handbook of child psychology* (pp. 793–828). Wiley. <https://doi.org/10.1002/9780470147658.chpsy0114>
- Chun, D., Smith, B., & Kern, R. (2016). Technology in Language Use, Language Teaching, and Language Learning. *The Modern Language Journal*, 100(S1), 64–80. <https://doi.org/10.1111/modl.12302>
- Elkins, K., & Chun, J. (2020). Can GPT-3 pass a writer's Turing Test? *Journal of Cultural Analytics*, 5(2). <https://doi.org/10.22148/001c.17212>
- Ferlazzo, L. (2023). ChatGPT: Teachers Weigh In on How to Manage the New AI Chatbot (Opinion). *Education Week*. <https://www.edweek.org/teaching-learning/opinion-chatgpt-teachers-weigh-in-on-how-to-manage-the-new-ai-chatbot/2023/01>
- Floridi, L., & Chiriatti, M. (2020). GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds and Machines*, 30(4), 681–694. <https://doi.org/10.1007/s11023-020-09548-1>
- Godwin-Jones, R. (2022). Partnering with AI: Intelligent writing assistance and instructed language learning. *Language Learning & Technology*, 26(2), 5–24. <https://doi.org/10.125/73474>
- Huang, K. (2023). Alarmed by A.I. Chatbots, Universities Start Revamping How They Teach. *The New York Times*. <https://www.nytimes.com/2023/01/16/technology/chatgpt-artificial-intelligence-universities.html>
- Hubbard, P. (1988). An integrated framework for CALL courseware evaluation. *CALICO Journal*, 6(2), 51–72.
- Hubbard, P. (2011). Evaluation of courseware and websites. In L. Ducate & N. Arnold (Eds.), *Present and future promises of CALL: From theory and research to new directions in foreign language teaching* (pp. 407–440). CALICO.
- Hubbard, P. (2021). *An Invitation to CALL*. Asia-Pacific Association for Computer-Assisted Language Learning (APACALL).
- Idder, A., & Coulaux, S. (2021). *Artificial intelligence in criminal justice: invasion or revolution?* International Bar Association. https://www.ibanet.org/dec-21-ai-criminal-justice#_edn9
- Koltovskaia, S. (2020). Student engagement with automated written corrective feedback (AWCF) provided by Grammarly: A multiple case study. *Assessing Writing*, 44. <https://doi.org/10.1016/j.asw.2020.100450>
- Krashen, S. D. (1982). *Principles and practice in second language acquisition*. Pergamon.
- Lorentzen, A., & Bonner, E. (2023). *Customizable ChatGPT AI chatbots for conversation practice*. *The FLTMAG*. <https://ftmag.com/customizable-chatgpt-ai-chatbots-for-conversation-practice/>

- McMurry, B. L., Williams, D. D., West, R. E., Anderson, N. J., Rich, P., J., & Hartshorn, K. J. (2016). An evaluation framework for CALL. *TESL-EJ*, 20(2).
- Mollman, S. (2023). ChatGPT passed a Wharton MBA exam and it's still in its infancy. One professor is sounding the alarm. *Fortune*. <https://fortune.com/2023/01/21/chatgpt-passed-wharton-mba-exam-one-professor-is-sounding-alarm-artificial-intelligence>
- Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464), 447–453. <https://doi.org/10.1126/science.aax2342>
- Paterson, K. (2022). Machine translation in higher education: Perceptions, policy, and pedagogy. *TESOL Journal*, 14(2). <https://doi.org/10.1002/tesj.690>
- Rosell-Aguilar, F. (2017). State of the app: A taxonomy and framework for evaluating language learning mobile applications. *CALICO Journal*, 34(2), 243–258. <https://doi.org/10.1558/cj.27623>
- Schmidt, R. W. (1990). The Role of Consciousness in Second Language Learning1. *Applied Linguistics*, 11(2), 129–158. <https://doi.org/10.1093/applin/11.2.129>
- The Douglas Fir Group. (2016). A transdisciplinary framework for SLA in a multilingual world. *The Modern Language Journal*, 100(S1), 19–47. <https://doi.org/10.1111/modl.12301>