

A ChatGPT-Based Contextual Vocabulary Trainer for English as a Second Language (ESL)

Sean Lydon Minton
Learning Design and Technology
University of Hawai'i at Mānoa
Mintons@hawaii.edu

Abstract

Many English language learners struggle to build the vocabulary and metacognitive strategies needed to adapt in new linguistic environments. This project addressed that challenge with Critical Reader, a custom ChatGPT-based instructional tool that helps learners develop vocabulary acquisition and retention habits.

The instruction begins with an introductory and “language hack” video, followed by guided practice using the Pre-Think, Assume, Use (PAU) cycle—a structured conversational routine. Grounded in Cognitive Load and Flexibility Theories (CLT & CFT), the GPT dynamically adjusts prompts based on learner level by adding context clues that support comprehension.

Evaluation data from four usability tests and 16 learner performance tests indicated the instruction met its objectives. Learners showed improved grammar and greater accuracy in inferring word meaning, while instructors reported faster vocabulary retention. Attitudinal survey responses reflected high engagement and perceived effectiveness.

Future improvements include API-based longitudinal learner performance tracking, expanded error analysis for personalized instruction, and multimodal features to support speaking and listening.

Analysis

Problem Statement: A Need for Direct Instruction

New English as a Second Language (ESL) learners at the Institute of Intensive English (IIE) must rapidly adopt new language learning strategies, including a simplified ‘critical reading’ approach for vocabulary acquisition and deployment. One impediment to learning this ‘critical reading’ approach is IIE’s rolling admissions policy, which redirects instructors’ attention to onboarding new learners rather than engaging deeply with language learning strategies. This shift restricts the instructors’ ability to provide personalized feedback, rendering the intervention only partially effective and potentially hindering learners’ development of related skills.

Learners frequently acknowledge that the ‘critical reading’ process aids integration. However, feedback from ad hoc focus groups over the past two years highlights the need for increased interaction with the instructor to deploy the ‘critical reading’ process effectively. IIE instructional staff regularly encounter the problems ‘critical reading’ addresses, and may also benefit from a familiarity with this process. Learners risk missing valuable language opportunities if they do not have foundational strategies for navigating unfamiliar linguistic settings.

Target Audience: Adult English as a Second Language (ESL) Learners

The primary audience with this problem are adult (18+) intermediate to advanced English learners. Relevant characteristics of this audience are their familiarity with mobile technologies, busy lives, and limited time for language learning. When practicing ESL, learners frequently make errors influenced by the structures of their first languages. In addition, they possess varying motivation levels, learning rationales and wide variations in ability, even within their assigned language level. These audience attributes increase instructors’ reliance on adaptive curriculum, which has proven challenging to create and deploy.

Despite these challenges, learners are dedicated to improving their English. They prefer interactive peer engagement over other instructional methods and favor games or activities over direct instruction. They require practical, effective, and easily implementable strategies that fit within their daily routines and interests.

To address the characteristics, needs, and challenges of learners in IIE classrooms (see [Appendix A](#)), any intervention must meet these specific criteria: it should be easy to implement by both the learner and instructor, use natural English contexts to facilitate implicit learning, adapt to both short bursts of microlearning and longer, focused sessions, apply to a broad range of learner-directed content, be designed for life-long use, and be easily adjusted to function across different English modalities.

Project Rationale

If learners cannot quickly adapt to effective self-directed learning processes like Critical Reading, they risk missing valuable language learning opportunities. A technology-based intervention might enhance feedback delivery and expedite the routinization of this cognitive language learning

exercise. In turn, this would allow improved exploration of novel contexts using learned material during class.

Implemented as proposed, this project may improve feedback given to learners, enabling them to accurately infer the meaning of unfamiliar vocabulary in new contexts while reading. Additionally, learners will report improved vocabulary acquisition, use, and retention. The first two weeks of the learner's journey may better emphasize 'jumping in' versus 'catching up,' which may lead to increased language production, confidence, motivation, and overall satisfaction.

The goal of this project is to design and evaluate original instruction that enhances feedback given to learners while learning and practicing vocabulary strategies in intermediate and advanced English as a Second Language (ESL) courses.

Literature Review

The Intervention: Critical Reading

The current version of the critical reading process guides learners through a structured approach to understanding new vocabulary within a given context. First, they skim the text to identify familiar words and concepts, creating mental connections with the vocabulary they already know. As they read more carefully, they mark unfamiliar words and use context clues, sentence structure, root words and affixes to guess meanings. This is followed by checking their guesses with a dictionary and finally, applying the new vocabulary in sentences to reinforce understanding. Throughout this cycle, learners actively build their language skills, using reflection and repetition to make new words part of their everyday language (see [Appendix B](#)). Learners report that this process is challenging, but very useful. They highlight the main bottleneck to adoption: interactions with, and feedback from the instructor.

The Problem: Feedback in SLA

Observations of performance and learner feedback illustrate the disproportionate impact of direct interaction on the outcomes of this intervention (see [Appendix C](#)). One-on-one tutoring consistently outperforms classroom instruction (Bloom, 1984; Nickow, 2020), highlighting the value of feedback in student learning. While direct tutoring may not be a viable solution for IIE's English learner demographic, exploring the characteristics of feedback that produce better learning outcomes provides valuable design insights. A meta-analysis on corrective feedback in Second Language Acquisition (SLA) suggested that the most effective feedback format involves short, computer-led instructional periods focusing on drill or activity-based tasks that allow for freely constructed responses (Li, 2010).

Statistical analysis of Li's data (2010) provided a detailed, codified understanding of feedback effectiveness in numerous ways. For instance, the data indicates that explicit feedback supports immediate language production, while implicit feedback enhances long-term retention (see [Appendix D](#)). Leveraging these findings requires connecting them to IIE's existing educational architecture to address the unmet needs of its learners.

The Audience: Adult Learners of a Second Language

IIE's organizational strategy aligns closely with andragogical theories, emphasizing a learner-centered approach that uses errors to enhance grammar knowledge with immediate real-world application (Kearsley & Culatta, 2018a). IIE also prioritizes engaging games and activities to deliver content effectively. It's crucial that this intervention integrates seamlessly into this structure while addressing key gaps in the cognitive and metacognitive aspects of language learning.

Observations of 'critical reading' in action show that while learners initially expect about 30% accuracy when guessing new word meanings from context alone, they achieve about 60% accuracy on their first attempt. This suggests that adults can expedite their learning by applying higher-order thinking strategies. However, these skills are often subconscious and initially challenging. Accounting for the intervention's relatively high cognitive load while leveraging the problem-solving abilities and skills it intends to facilitate is important for its success.

Cognitive Load Theory (CLT) suggests several enhancements, some as simple as replacing inconsistent verbal instructions with consistent video content. Conversely, more complex strategies might integrate expert 'schemas' into the content's scaffolding and use adaptive feedback to modify task difficulty, highlight key clues, maintain user interest, or introduce variability (Sweller, 1988, 2020). These simple considerations of the learner's attention or complex real-time adjustments would reduce cognitive load while remaining adaptable to the diverse learning needs of IIE classrooms.

Cognitive Flexibility Theory (CFT) provides further insight into unstructured learning environments that mirror real-world experiences. It posits that repeated exposure to learned material in varying contexts strengthens learner "processing skills" and builds robust "knowledge structures" (Spiro et al., 1991). Instructor observations show that learners most familiar with 'critical reading' demonstrate improved spontaneous application of these skills and structures in real-world scenarios. CFT's adherence to numerous modalities, interconnected content domains and contextual knowledge construction (Kearsley & Culatta, 2018b) better connect this intervention to its overarching goals of learner integration.

These theoretical insights can help ensure the intervention retains its valuable features while addressing key barriers to adoption. Ideally, this intervention systematically trains learners to adaptively "think about their own thinking" and enhances its power as a metacognitive link (Kim & Chang, 2008) to IIE's broader immersive learning architecture (see [Appendix E](#)). With the characteristics and the theoretical basis of effective feedback defined, we now examine optimal delivery methods.

The Delivery: Adaptive Feedback

Research has shown that Intelligent Tutoring Systems (ITS) improve student learning across various domains (MacLellan & Koedinger, 2020). They have been shown to be most effective in language learning (Tlili et al., 2023), even though they may not perform as well as human tutors (Steenbergen-Hu, S. 2014). To avoid ineffective designs, any ITS should intentionally constrain

subject matter, reduce learner barriers, consider the learner in detail, and function dynamically (Nwana, H., 1990).

The learning material in Critical Reading is simple enough to be completed using short answers or multiple choice responses. However, it would benefit from being automatically scaffolded by difficulty, and randomized for variety according to the theories and research listed above. A clear connection can be drawn between these three bodies of research: an ideal approach provides empirically supported effective feedback, supports broader attempts at real-world integration, and is well scaffolded to decrease cognitive load.

The ability to automatically design and scaffold feedback requires the configuration of three models, or functional analogies of them; the tutor, the learner and the content or expert. (Nwana, H., 1990), for further examination in the Design and Development sections of this document.

Design

Overview

The goal of this project is to design and evaluate original instruction that introduces learners to the Critical Reading process and allows them to practice it. Learners will gain confidence while acquiring and using new vocabulary. This intervention consists of two videos and a generative AI platform based practice tool.

Scope and Sequence

The Critical Reading process contains multiple steps and numerous potential topics (see [Appendix E](#)). Three findings in learner feedback on this process were critical in reducing the project's scope:

1. Learners believe, and the Instructional Designer agrees, that this process can be learned actively.
2. Learners apply this process to a wide range of English modalities, and the intervention should consider those use cases.
3. Learners can accurately guess new vocabulary at a success rate of approximately 60% on their first attempt.

These findings prompted changes in the scope and potential sequencing of the lessons. The first observation illustrates that direct feedback is critical to success in this process. The intervention can remove some specific topics that relate only to reading, as they won't be relevant in other language modalities, and they can be quickly covered during class time as necessary. Third, the intervention eschews direct instruction in many 'clues' and 'methods.' The learner's initial rate of guessing new vocabulary provides evidence that they do not need to be taught explicitly. Instead of teaching 'context clues' directly, the intervention uses them to design the prompts that learners receive.

Revised versions consist of two videos and a practice tool. First, an introduction provides important information on how the exercise connects to courses at IIE, a brief summary of the process, and information on using the practice tool. The second video contains tips and tricks, with a summary of the science behind the process. The videos are presented to the learner prior to beginning practice using the generative AI tool (see [Appendix G](#)).

Strategies and Method

This intervention uses microlearning strategies in a direct-instruction video coupled with experiential tasks using an indirect or experiential practice platform. The process is divided into three stages: Pre-think, Assume, and Use. Each stage contains different tasks for the learner to complete. In the performance of these tasks, the practice tool will produce prompts, and the learner will respond (see [Appendix H](#)). These tasks are the performance objectives of the intervention, which the GPT will evaluate and provide feedback on.

Performance Objectives

Performance objectives were formulated to guide the development of the intervention and its evaluation methods. Prior versions of the Critical Reading process only tested the learner's ability to guess new vocabulary, an 'Assumption' task. With a technology-based intervention, especially a generative AI platform, it is possible to evaluate other points in the learner's journey. With that in mind, the evaluation was expanded to four major tasks: Connection, Assumption, Rationale and Sample tasks (see [Appendix I](#)).

Technical Details

Learner feedback on this process, alongside other considerations of the target audience, suggested that an Android and iOS capable platform with the ability to deliver video, allow learner inputs, and deliver language use feedback would best meet learner needs. Furthermore, the types of prompts that the learner needs to practice critical reading skills can be delivered consistently by a custom ChatGPT, hereafter referred to as the Generative Pretrained Transformer (GPT).

An AI video production and editing platform, InVideo, created introduction videos. The GPT delivers those videos, hosted on YouTube, in its first interaction with the learner. The GPT then walks the learner through rounds of practice. This method required replicating the in-class procedures enacted by the instructor, by creating learner, expert, and instructor models to act as the ChatGPT's instructions:

1. Learner model: Contains target vocabulary and grammar conventions by Common European Framework of Reference for languages (CEFR) level (EnglishProfile.org, 2006), which informs the instructor model and gets more detailed as the learner progresses.
2. Expert model: Contains rubrics used to assess the learner. This model allows the instructor to continually adapt to learner skill by adjusting the difficulty of prompts, attempts to improve the learners' assessment scores, and provides details to the learner model.

3. Instructor model: Contains learner-instructor interaction flowcharts, performance objectives, and other instructions that appropriately prompt the learner at the right moment, keep the learner on track, and offer feedback.

Other programs used in the production of materials included Google suite applications like Slides, Forms and Sheets. QR codes delivered the GPT link, when necessary. When the learner wants to review their scores, the GPT produces an assessment score report (see [Appendix P](#)), and a link to a Google Form that accepts the score report for further analysis.

While custom ChatGPT's can deliver materials in an expected manner, learners are currently restricted to free tokens unless they are paid ChatGPT users. Use of this platform, as it currently stands, required workarounds for usability studies and reliable in-depth analysis. Further details on these score reports and testing format workarounds can be found next, in the methodology and evaluation section of this intervention.

Methodology

Goal

The purpose of this evaluation was to assess the usability and learning effectiveness of the ChatGPT-based Contextual Vocabulary Trainer. Specifically, the evaluation aimed to determine whether learners can understand the videos the GPT provides, believe and demonstrate that the GPT will help them with their language skills, and that the Critical Reading process teaches productive learning habits.

Participants

Participants included language students at levels B1 to C1 on the Common European Framework of Reference for Languages (CEFR) and language instructors from the Institute of Intensive English (IIE) in Honolulu, Hawaii. Participants falling into three general categories tested the tools while using a paid version of ChatGPT (see [Appendix J](#)). It is important to note that the GPT's performance according to its specifications will dictate learning outcomes. As such, the GPT is also considered a usability participant.

Processes, Instruments, and Data Analysis

Evaluation utilized three data collection domains. Each domain employed specific processes and instruments to collect and analyze data. The majority of tests were conducted in a classroom setting, led by the instructional designer

1. Usability tests collected opinion surveys and GPT usability data in tallied rubrics and notes from dynamic, one-on-one conversations with learners.
2. Opinion surveys collected the bulk of demographic and attitudinal data.
3. The GPT delivered learner performance data, which was collected in Google Forms.

Usability Data

Usability pre-tests, conducted early in the GPT's development, informed precise data points for future collection (see [Appendix K](#)). Standardized usability test sessions involved the instructional designer scheduling one-hour, in-person tests with three learners and one English instructor.

During each session, learners accessed a paid GPT account on a laptop, watched the introductory videos, completed approximately three PAU (Pre-Think, Assume, Use) loops, received a baseline score report, worked through additional loops, received a final score, and completed an opinion survey. Meanwhile, the instructional designer monitored GPT behavior against specifications and held follow-up conversations regarding usability findings. Data collection methods included a tallied rubric in (see [Appendix L](#)), and qualitative notes from tailored conversations with the learner (see [Appendix M](#)). Analysis of the GPT transcripts with the learner highlighted deviations from the designers expectations and then determined adjustments in the GPT's configuration to better align with learner needs.

Learner Performance Data

All tests were performed using a paid GPT account, where learners had sufficient time with the tool to receive reliable scores. While interacting with the GPT, learners were assessed on performance objectives in Connection, Assumption, Rationale, and Sample tasks. The GPT evaluated these based on the Novelty, Accuracy, and Relevance of the learners' responses (see [Appendix O](#)). The learners then instructed the GPT to produce reports, which summarized the total number of completed tasks, scores, common mistakes, and positive language samples (see [Appendix P](#)). Learners submitted these reports via Google Forms, where the data was analyzed in Google Sheets. The primary focus of performance data analysis was the difference between baseline and final scores.

Attitudinal Data

All participants were given access to the GPT and completed an opinion survey in Google forms, delivered via the GPT or QR code. The Google Forms opinion survey used logic functions to separate responses by various participant characteristics and gathered demographic, usability, and attitudinal data (see [Appendix N](#)). The survey covered three broad categories:

1. Learners' ability to understand the videos provided by the GPT.
2. Learners' opinions on the GPT's outputs.
3. Learners' opinions on the Critical Reading process in general.

The designers primary objective was to ensure that experienced learners agreed the critical components of the original version were present in the new tool and rated it as generally superior. Beyond that confirmation, comparing results from experienced and new learners would yield rich data for design refinements. For instance, if positive feedback from experienced learners did not match data from new learners, it would indicate specific problems with GPT prompts, the introduction video, or other components of the GPTs back end configuration. More specifically, the

highest priority fixes may be defined as those with the greatest value difference between new and experienced learners.

Results

Test Demographics

Among the 20 usability and learning performance participants, 53% spoke an East Asian language, 37% spoke a Romance or European Language while 11% spoke a Middle Eastern or South Asian Language. Language proficiency was evenly distributed, with a focus on intermediate levels ([Appendix X](#)), ensuring a representative sample of IIE learner demographics.

Usability Test Result

Usability studies evaluated part of the intervention's effectiveness by analyzing learner experience. Four participants watched the videos, used the GPT and completed an opinion survey regarding video comprehension, the GPT's outputs, and the overall Critical Reading process. Meanwhile, the instructional designer observed GPT-learner interactions and recorded the GPT's responses for analysis. Post-interaction discussions between the designer and the learner focused on areas where the GPT's configuration did not align with learner needs.

Data collected using a Likert scale ranging from one (strongly disagree) to five (strongly agree), indicated moderate agreement in four important categories: the 'understandability' of the introduction videos ($M=4.25$, $SD=0.96$), the video's explanation of how to use the GPT ($M= 3.50$, $SD=0.58$), current features of the intervention ($M= 4.00$ $SD=1.00$), and the overall improvement of the intervention over previous versions ($M = 3.00$, $SD=1.00$). Complete details of these results can be seen in [Appendix R](#)- Table R1.

GPT performance data provided deeper insight into the intervention's effectiveness, by codifying and reviewing the GPT's outputs with learners. Amongst 55 total GPT-learner interactions across four usability tests, 26 GPT outputs (47%) met designer expectations, while 22 GPT outputs (40%) exhibited known errors. Three GPT outputs (6%) were hallucinations (undesirable and unexpected responses), while four GPT outputs (7%) were novel (desirable and unexpected responses) (see [Appendix S](#)). Post-test discussions revealed that of the 26 GPT outputs meeting expectations, 23 (89%) required no modifications, while the designer and learner agreed that three (12%) of those outputs required adjustment. Out of the 22 GPT outputs demonstrating known errors, 13 (59%) required modifications while the learner and designer agreed to eschew modifications to nine (41%) of those outputs.

This data led to several modifications to the videos. The GPT was given a new 'action' to consistently retrieve the second instructional video, which provides additional details on the Critical Reading process. Given the extensive revisions to the video content and structure, future versions will be translated into the learners primary language.

Three key changes were made to the GPT's configuration. First, the GPT now consistently addresses learner errors. Early versions sought to reduce cognitive load by omitting some corrections, but testers emphasized the value of even minimal feedback. As one participant noted "Even simple feedback is helpful. If it points out mistakes, in any way... It's good. You can always ask the GPT for further explanation." Second, the GPT better prioritizes providing sample sentences instead of direct explanations of new vocabulary. Agreeing that the Critical Reading process will form productive learning habits, one English instructor test participant highlighted that "filling in knowledge gaps [via Critical Reading process] is an efficient method for long-term vocabulary memorization." Third, modifications to the GPT's scoring system ensured it accurately represents learners' language abilities rather than their reproduction of GPT-generated phrasing. For a complete list of modifications, refer to the change log in [Appendix K](#).

Learner Performance Results

Learning performance tests measured 16 learners' proficiency across four task types (Connection, Assumption, Rationale, and Sample tasks) and three evaluation dimensions (Novelty, Accuracy and Relevance). The GPT recorded total tasks completed, assigned scores from one to ten for each task, and calculated an average of the last three scores in each category to track progress.

Results provided insights into learner productivity at different proficiency levels, and predictable trends emerged. For example, A2 through B1 learners created connections at a higher rate and were more limited in their ability to produce rationales than advanced levels, while those advanced levels performed more tasks in the baseline score, completed a wider range of tasks, and demonstrated slightly greater overall productivity ([Appendix T](#)).

Performance data analysis indicated modest gains across tasks and dimensions, as seen in [Appendix U](#). Instructors used GPT's analysis capabilities to define behaviors that contributed to increased scores. For instance, B1+ level learners scored significantly higher in their Connection scores. The GPT highlighted grammar structures and cognitive processes that contributed to that increase. ([Appendix V](#)- Sample V1). These recommendations, verified by IIE instructors, give highly practical suggestions to increase fluency. Instructors confirmed that the GPT's insights, particularly the mastery of subordinate and relative clauses, were particularly astute. Similarly, increased Accuracy scores among A2–B1 learners corresponded with improvements in subject-verb agreement and word order, while also mentioning that A2 learners required additional grammar support to apply conventions consistently in new contexts ([Appendix V](#)- Sample V2).

Attitudinal Survey Results

Data collected during usability testing informed GPT configuration and video modifications ([Appendix R](#)). The same data was collected again during learner performance testing. Modifications to the GPT's configuration and the videos had a positive effect on those metrics, where minor variations in average scores are attributed to differences in sample sizes. Overall, learners had positive perspectives on both the Critical Reading process and the GPT's impact on their language learning ([Appendix W](#)).

This intervention is integrated within IIE's broader pedagogical architecture, making it essential to assess whether learners with Critical Reading experience recognized key instructional elements and whether new learners found the GPT tool beneficial. Experienced learners agreed that the GPT retained essential components while improving upon previous versions. New learners strongly agreed that the intervention would enhance vocabulary acquisition, language retention and engagement (see [Appendix W](#)- Table W2).

To further interpret these findings, qualitative comments were categorized and analyzed (see [Appendix Y](#)). Among learners who provided feedback, 28% wanted a more diverse range of learning tasks and styles of user interface, 23% expressed appreciation for the GPT-provided grammar assistance, and 18% believed direct instruction remained necessary for effective learning. Notably, 9% of participants observed that the GPT lacked direct instruction. This comment was popular among language instructors and instructional designers, and underscore the necessity of further refinement to balance GPT-driven learning with structured instructional support.

Discussion

Students reported that changes made after usability testing significantly improved the intervention's effectiveness. While additional adjustments and multilingual video options are still necessary, most learners understood the instructions well enough to continue using the tool independently. This aligns with experienced learners' beliefs that the Critical Reading process is best learned through action, but posed a burdensome cognitive load for new learners. Additionally, videos must be adapted for use with and without the GPT tool to account for other variables affecting curriculum delivery at IIE.

The GPT generally followed its configured instructions, and learners found its feedback helpful despite occasional imperfections. Notably, there was a positive shift in Assumption task Accuracy scores, suggesting that the GPT effectively improved learners' ability to infer word meanings from context—the primary objective of the Critical Reading process. This improvement likely stems from the intuitive nature of the Pre-think, Assume, Use (PAU) cycle.

Reduced potential for embarrassment during this low-stakes exercise provides more opportunities to utilize productive failure, and the immediate feedback loop strengthens new cognitive pathways while automated scaffolding from implicit to explicit corrections reduces cognitive load. The GPT's provision of varied sentence structures allows learners' to begin noticing semantic and syntactic patterns, increasing their inferential accuracy. As learners become consciously aware of this process and experience early success, it reinforces the strategy's perceived utility, encouraging continued use across varied contexts and developing productive language learning habits.

Learners thought the GPT was a useful tool but noted that it would benefit from a more engaging user interface and built-in direct instruction for grammar or vocabulary topics. These responses align with expectations for this learner demographic and reflect current limitations of the ChatGPT platform. Further information on how this intervention participates in moving IIE's current state to its desired state can be found in [Appendix Z](#). In particular, the GPT needs more specific error classification protocols to trigger consistent and accurate grammar explanations. Although effective

in its current form, the intervention remains limited by its text-based interface. Incorporating multimodal features for speaking and listening functions, and a wider variety of instructional strategies, could significantly enhance engagement and learning outcomes.

Conclusion

The Critical Reading process itself was designed to help learners navigate unfamiliar linguistic environments. The intervention aimed to enhance feedback during this process. Results showed that learners inferred unfamiliar vocabulary more accurately after using the GPT, suggesting that the intervention met its instructional goal. Although more longitudinal data is needed, learners reported gains in vocabulary acquisition, use, and retention, and experienced more efficient and consistent feedback during this challenging metacognitive exercise. One early usability tester, an instructional design professor specializing in AI, noted: “[This tool is] a useful supplement to systematic learning experiences, for students who already have some fluency in the target language.”

This statement illustrates the tool's strengths and limitations. Learners valued the random, student-led interactions and believed that intuitive understanding of vocabulary aids acquisition. However, they also noted the lack of direct grammar instruction. Addressing this issue may require further fine-tuning and integration. Although the tool's learning impact is evident, long-term data would offer deeper insights. Future development should address these needs through API integration, improved UI, and expanded instructional capabilities.

The Critical Reader GPT now serves as a framework for broader linguistic applications. The Novelty, Accuracy, and Relevance (NAR) rubric will continue to evolve, informing learner models for other GPT-based tools. The Critical Reader framework has already been implemented in Spanish and Farsi versions and other components of the GPT's configuration can be adapted to diverse instructional objectives across different applications in a modular, scalable format.

This project underscored the value of pre-usability testing during development, and the iterative design process. Instructional design methods in feasibility and basic factor analysis, and the designer's early and close proximity to the test participants were essential for a detailed understanding of the problem. All of these helped ensure close alignment with both organizational goals and user needs. In addition to replicating some of the other concepts in this document, intelligent tutoring system designers should prioritize identifying the interacting variables that contribute to learners' misunderstandings. This is especially important when access to learners, or robust learner model data is limited.

Ultimately, this project demonstrates that when grounded in theory and tested iteratively, AI tools like ChatGPT can extend the reach of effective instruction, and Critical Reader offers a replicable model for AI-enhanced learning that balances innovation with educational rigor, and technical capability with usability.

Appendices

A. Appendix A

[\(Back\)](#)

Language Groups at the Institute of Intensive English

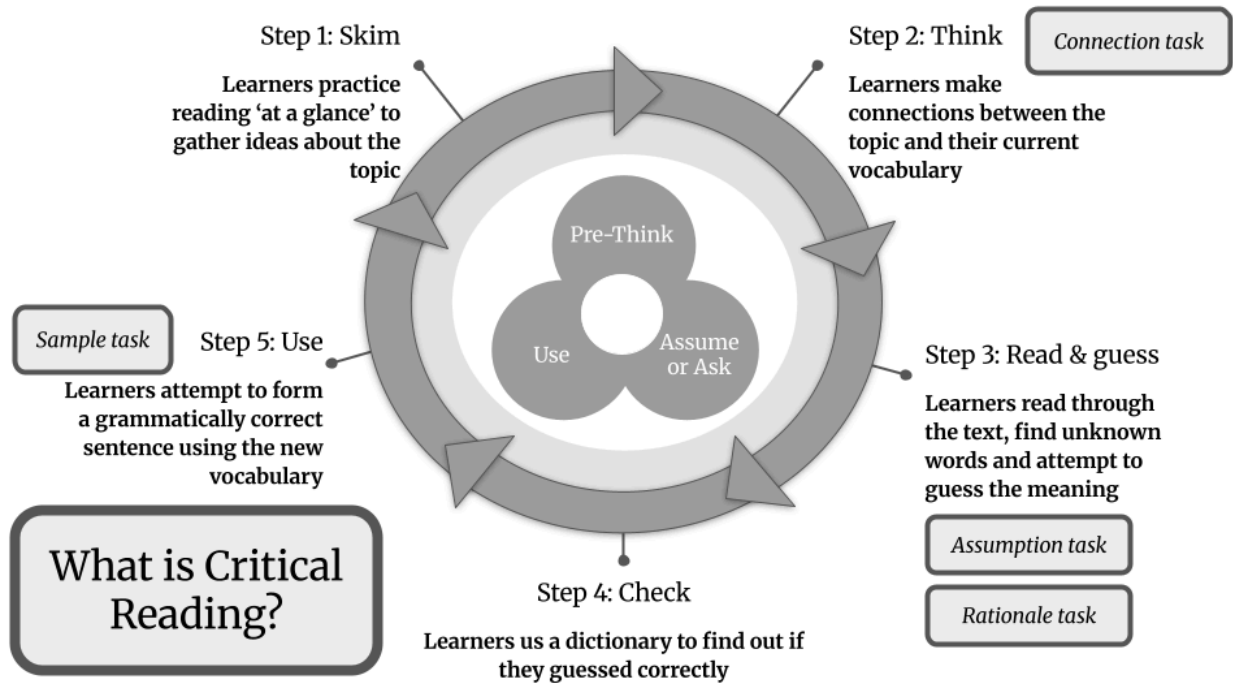
- This audience analysis considers the following courses at the Institute of Intensive English:
 - Courses related the Common European Framework (CEFR) levels and general tutoring classes
 - Excludes courses in specific focuses
- East Asian language groups include Japanese, Korean and Chinese
- European language groups include Germanic and Romance languages

Cognitive	Affective
<ul style="list-style-type: none"> • The school ranges from beginner to advanced levels; across all domains (reading, writing, listening and speaking) • All students exhibit English mistakes derived from grammar construction of their primary language • Varied exposure to different styles of instruction 	<ul style="list-style-type: none"> • Prefer to learn through games and activities over direct instruction and peer-to-peer interaction vs. other instructional modalities • Getting a mid to long term US visa is easier than other methods. <ul style="list-style-type: none"> ◦ ~20% of students are highly motivated to learn English, as measured by qualitative indicators of intellectual curiosity about the language itself • Report often being too busy to focus on language
Physiological	Social
<ul style="list-style-type: none"> • Characteristic of general population • Have ample experience using mobile technologies in the classroom 	<ul style="list-style-type: none"> • Characteristic of general population • In a new place with varying degrees of ability to rely on their primary language • Range of different nationalities, ethnicities and races • Generally affluent <ul style="list-style-type: none"> ◦ Younger ages skew to parental income ◦ Older ages skew to high-earners

Appendix A: Characteristics of English Learners at IIE

B. Appendix B

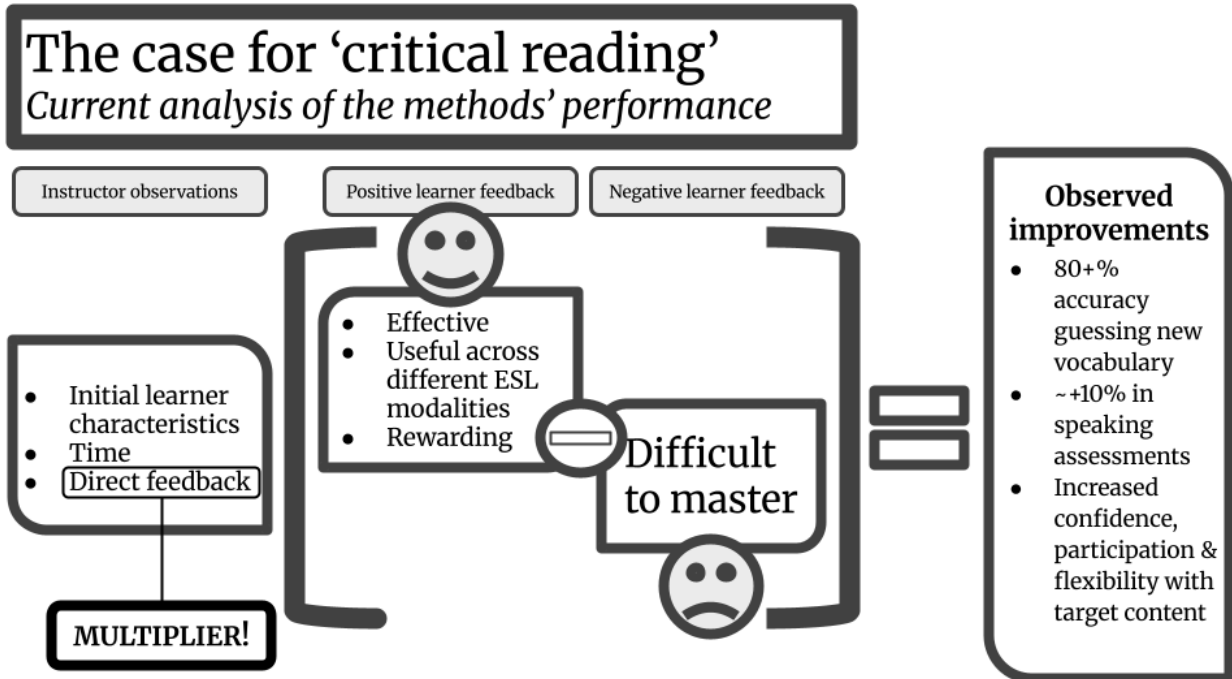
[\(Back\)](#)



Appendix B: A brief explanation of the original PAU loop

C. Appendix C

[\(Back\)](#)

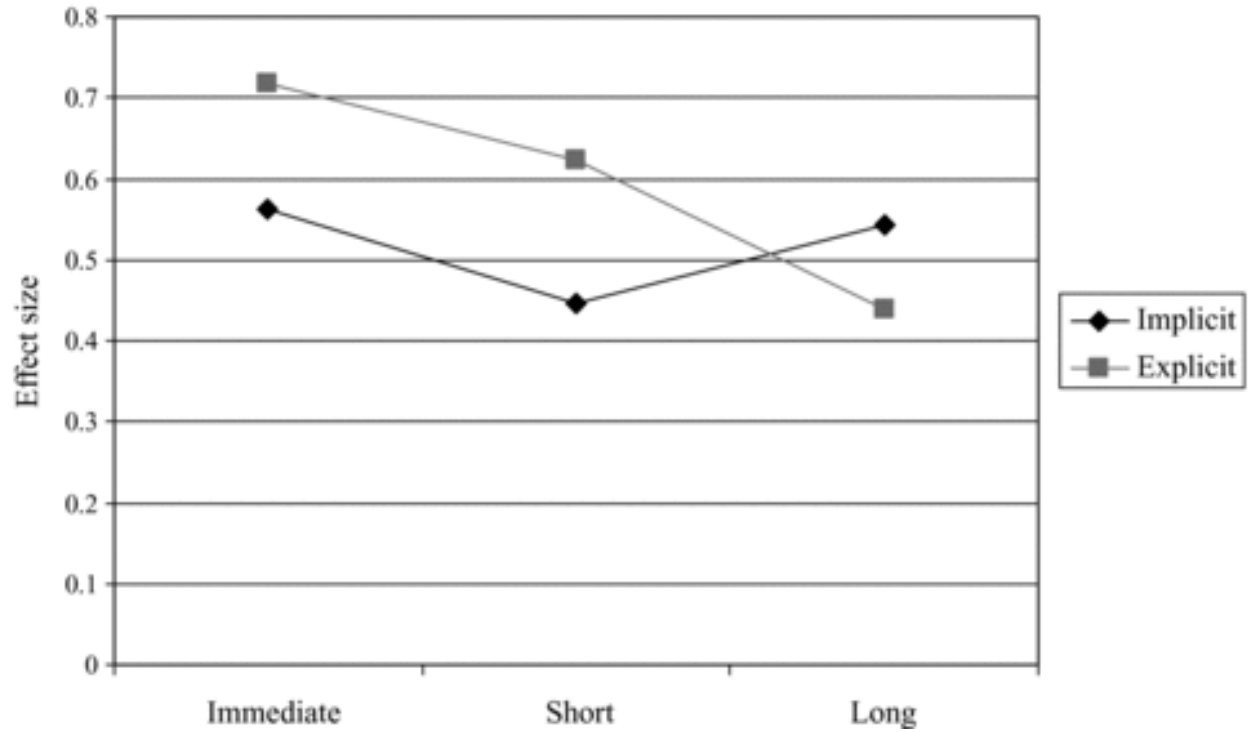


Appendix C: Summary of the process's current outcomes

D. Appendix D

[\(Back\)](#)

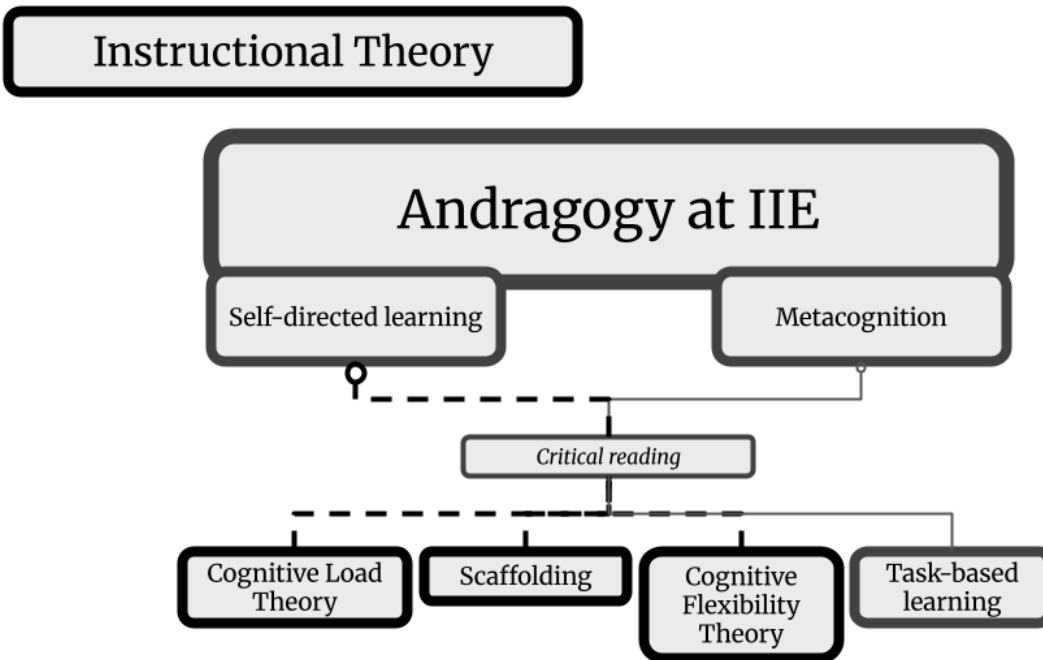
This data indicates that explicit feedback supports immediate language production, while implicit feedback enhances long-term vocabulary retention.



Appendix D: "Implicit and explicit feedback: Change of mean effect size over time." (Li, 2010)

E. Appendix E

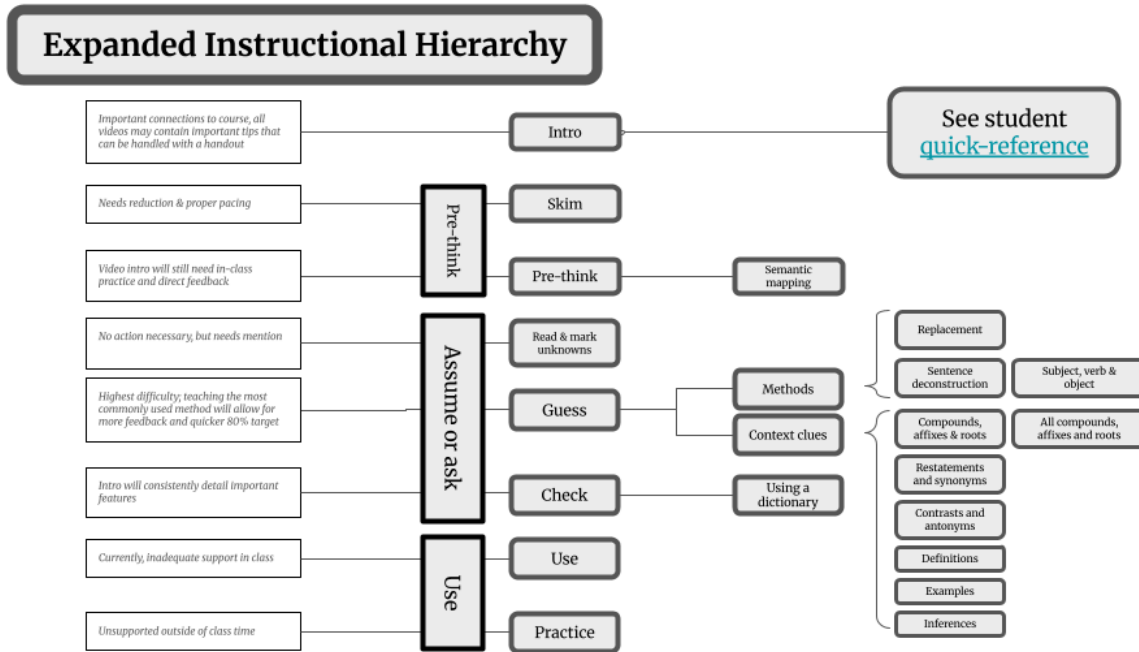
[\(Back\)](#)



Appendix E: Overview of theories and pedagogies at IIE

F. Appendix F

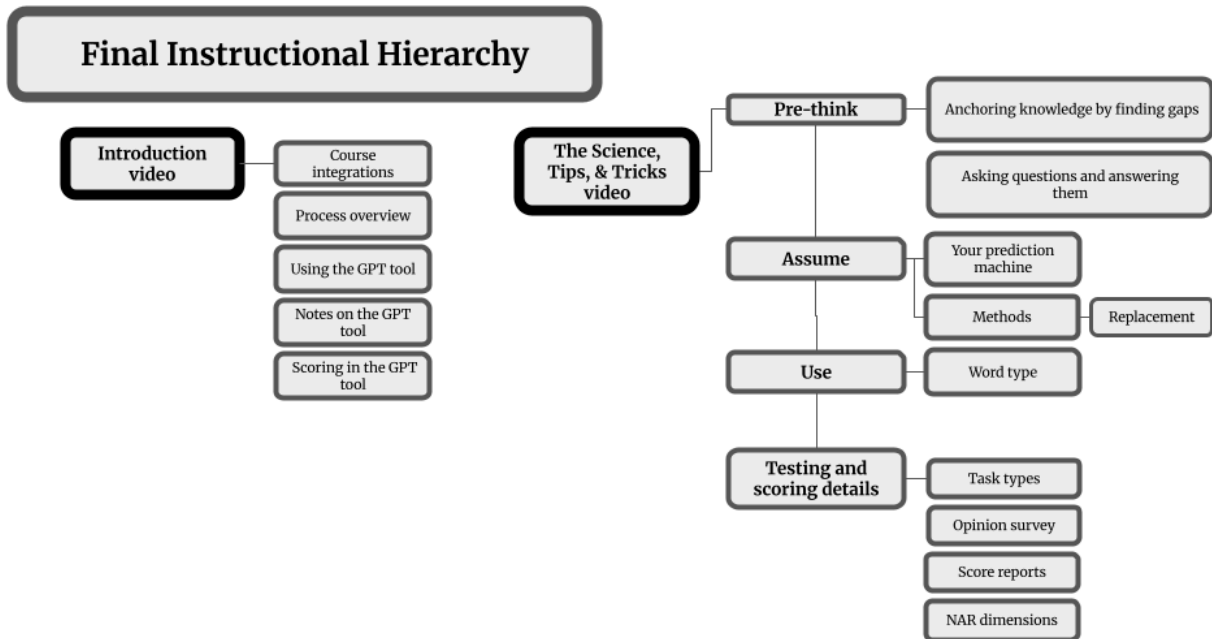
[\(Back\)](#)



Appendix F: Overview of skills and stages in the critical reading process and valuations of previous versions.

G. Appendix G

[\(Back\)](#)



Appendix G: Revised scope and sequence of instruction

H. Appendix H

[\(Back\)](#)

Prompt Design Document Sample

Below is a sample prompt for a C1 level student from the primary document directing the GPT in matching the learners language level and providing appropriate prompts. This document can be found here in full detail, [Prompt Design Document](#).

Pre-thinking stage prompts defined	Assumption stage prompts defined	Continuation suggestions	Clue suggestions	Rationale suggestions
Imagine you're in a bustling open-air market in a foreign country. There are vendors selling fresh produce, spices, and handcrafted goods. You overhear some people haggling over prices and see others sampling food.	As you approach a vendor with a colorful display of interesting items, you hear him say "You won't believe the provenance of our artisanal jewelry, everything is antique!"	"You like that one? Notice the patina on this leather. Good leather needs to be used! I don't know, in my opinion... well cared for, used leather looks better than anything else."	For provenance, mention where items are from or something about its history. For artisanal, mention that it is made in small batches by hand using traditional methods or similar. For patina, mention that [the vendor strokes the surface of the leather] or [points out the slight mistakes on the jewelry's surface]	Provenance sounds like provincial, province or like a place, artisanal sounds like art, patina sounds like paint or something about a surface since the examples refer to the way something looks.

I. Appendix I

[\(Back\)](#)

Performance Objectives

Definitions

Known vocabulary	Vocabulary that the learner can be reasonably expected to understand according to their CEFR level
Target vocabulary	Vocabulary that the learner is expected to learn and use according to their CEFR level

Key

<i>Task type & objective</i>	<i>Condition</i>	<i>Audience, behavior & degree</i>
----------------------------------	------------------	--

Introduction

General task	After completing an introduction to the process	Language learner will use the practice tool to successfully navigate through next stages as measured by total tasks completed
General task	After completing the entire critical reading process	Experienced language learners with prior experience in the process will express a positive opinion about the intervention
General task	After completing the entire critical reading process	All language learners will provide opinion data on the intervention's effectiveness via an opinion survey

Pre-thinking

Connection	When presented with a prompt containing known vocabulary (IE one sample newspaper headline)	Language learner will write a brief (~1 word to 2 sentence) connection to the content that meets the criteria of the rubric [+/- % of average]
------------	---	--

Assume or Ask

Assumption	When presented with the second part of the prompt containing target vocabulary (IE one corresponding body paragraph from the previous article)	Language learner will write an a brief (~1 word to 2 sentence) assumption about target vocabulary's definition, meeting rubric criteria [+/- % of average]
Rationale	When the GPT asks if the learner has a reason for their guess	Language learner will write a brief (~1 word to 2 sentence) rationale, meeting rubric criteria [+/- % of average]
General task	When the learner has asked & been introduced to their desired unfamiliar vocabulary	Language learner will successfully navigate to next stage as measured by total tasks completed

Use

Connection	When presented with the second part of the prompt containing known vocabulary (IE one corresponding body paragraph from the previous article)	Language learner will write a brief (~1 word to 2 sentence) connection, meeting rubric criteria [+/- % of average]
Sample	After learning a new phrase, sentence or word	Language learner will write a brief (~2 word to 2 sentence) sample using the new phrase, sentence or word, meeting rubric criteria [+/- % of average]
General task	When the learner has tested their use of the desired unfamiliar vocabulary	Language learner will repeat the process until they receive an average of 7.5 out of 10 points for Accurate Assumption scores over 3 most recent rounds of the PAU loop.

J. Appendix J

[\(Back\)](#)

Test Audience and Format

Usability test (n=4)		
Learning Performance (n=16)		
<i>Experienced learners</i>	<i>Language learners</i>	<i>Instructors</i>
Language learners who have experience with Critical Reading. It is essential that this group agree that new versions include the same important content that will assist inexperienced learners.	Learners with no experience in Critical Reading. This groups' data will provide an alternative view to the data from the experienced learner group.	Primarily consisting of instructors from IIE, the data from these tests provide necessary weight to the opinions of language learners.

Usability test (n=4)
<i>The GPT's outputs</i>
It is critical to observe and measure the interactions between the instructor and the learner to ensure that the GPT is appropriately calibrated to deliver effective, accurate and appropriate prompts, feedback, grammar assistance and assessment scores.

Usability test (n=4)
Learning Performance (n=16)
<i>Long format test</i>
Learners used a paid GPT account for better access to the tool. All usability studies are completed in the long format. These tests were the preferred format, as they contained the most accurate learner performance data and the transcripts provided valuable fine-tuning material for future versions of the GPT.

K. Appendix K

[\(Back\)](#)

Change log

A complete list of major changes made to the Critical Reader's configuration.

Contributors: Various language learners levels B1-C1, including learners of languages other than English (n≈40), IIE Language Instructors (n≈5), Instructional Designers (n≈5)

Date	Results	Modifications
Pre-test 11/1/2024 (Mandarin version)	GPT was not correctly following the PAU loop	Change process/instruction list to a flow chart
Pre-test (Mandarin version)	Relationship between 'Pre-think' stage prompts and 'Assume' stage was inconsistent	Develop prompt guidelines, no fine tuning in this model
Pre-test 11/1/2024 (Mandarin version)	GPT did not adequately adjust to user level	Incorporate grammar profile datasets by CEFR level and ensure CEFR is tied to other self-assessments of language level & add 'write as much as you can' & 'tell the GPT you know all of that vocabulary' tips to intro video. Dataset to Bibliography.
Pre-test 11/4/2024 (English v1)	Must instruct users to DL and register for ChatGPT	Users will likely do this when they receive QR, might need to be prompted
Pre-test 11/4/2024 (English v1)	No custom GPT use with GPT v3.5. Tokens limited to ~1 PAU loop	Workaround in GPT Teams or using ID'er account for testing
Pre-test 11/4/2024 (English v1)	GPT offered vocabulary explanations prior to provoking an assumption about target vocabulary	Refine flow chart in 'Assumption' task section
Pre-test 11/4/2024 (English v1)	Intro video needs restructure, slow, verbiage changes, scene change clarity, visual design.	Review interactive videos on other platforms. Otherwise restructure in InVideo

Pre-test 11/4/2024 (English v1)	GPT was incorrectly identifying which tasks had been completed	Revise instructions in flow chart, ensure clarity with 'only if-then' statements. Arrow direction may be affecting flow.
Pre-test 11/4/2024 (English v1)	Users agreed that the GPT's tone should match the instructors style	Begin fine tuning in 'sass.' Long format, use edited long form conversations as samples
Pre-test 11/4/2024 (English v1)	Users suggested that the GPT train different word forms and topic	Add 'you can instruct the GPT to do what you want' to intro video with other tips
Pre-test 11/4/2024 (English v1)	Assessments were produced in inconsistent formats	Revise assessment protocols in training; test Google Scripts to parse data
Pre-test 11/7/2024 (English v1)	Instructor test noted that it was 'a lot of words'	Appropriate for testing level; C1. Test at beginner stages for GPT's simplification or reversion to primary language. Prompts will still be too wordy at other levels.
Pre-test 11/7/2024 (English v1)	Instructor attempted to create multiple users in one GPT and the GPT could not differentiate	This may not occur often, may mention in intro video that each GPT is personalized to one user
Pre-test 11/7/2024 (English v1)	Instructor mentioned that feedback may not be detailed enough	Mistakes were primarily in spelling and punctuation, which the GPT was specifically told to ignore; may delineate by level what types of feedback are appropriate
Pre-test 11/7/2024 (English v1)	Instructor had questions about how colloquial the language would be, and gave suggestions	May add corpora of different vocabulary modalities (IE formal vs informal) (See COCA corpora)
Pre-test 11/7/2024 (English v1)	Instructor suggested that more than one guess might be good	Fine tuning & change in flow chart; 'Assumption' stage
Pre-test 11/10/2024 (Mandarin v2)	Weighted scoring by task may force inaccurate evaluations	Revise weighted scoring in instructor directions
Pre-test 11/12/2024 (English v2)	GPT was inserting target vocabulary into the pre-thinking stage (IE assumption tasks in pre-thinking)	Add performance objectives by stage for clarity into new flowchart
Pre-test 11/12/2024 (English v2)	Designer suspects prompts can be overly descriptive	Test in further usability studies. Fine-tuning for focus on grammar and additional information in prompt design (a

		prompt output template may provide better guidance and keep GPT on task)
Pre-test 11/12/2024 (English v2)	Reconsider “total tasks” vs “total loops” for assessment and check that task instances are correctly counted	Revise NAR rubric, Score report guidelines, and flow chart.
Pre-test 11/12/2024 (English v2)	Revise weighting scores by task	May not be necessary and difficult to check for accurate calculations.
Pre-test 11/12/2024 (English v2)	Learner requested scoring method mention in video	Add short blurb to Intro on InVideo, reload to YouTube and relink in GPT
Pre-test 11/12/2024 (English v2)	Learners in B2- C1 mentioned that video pacing was good, but may be less comprehensible at lower levels	Test with B1 and below if possible, revise verbiage in video, or create versions in other languages. Calibrate by testing in other languages
Pre-test 12/12/2024 (English v2)	Learners suggested screenshots to guide students in completing the assessments	Test in further usability students, with remote students if possible. Add opinion survey to GPT
Pre-test 12/14/2024 (Portuguese v2)	GPT was not delivering visual prompts as directed for very early learners.	Revise across flow charts, instructions and prompt design.
Pre-test 12/14/2024 (Portuguese v2)	Updates hard to manage across languages	Remove all language specific information to the top section of the instructor directions document for paste into the GPT header. All other information to the main body for training input.
Pre-test 12/17/2024 (English v2)	Rearranging GPT main body and instructor directions cause catastrophic failure	Revert to previous versions. All adjustment instructions, document summaries and language specific information need to be carried in GPT main body
Pre-test 12/17/2024 (English v3)	May need custom privacy policy for storage of learner data	No need if covered by Google public use, statement of IDer intent though.
Pre-test 12/17/2024 (English v3)	Double check that GPT is counting tasks correctly, esp. whether or not it is correctly counting tasks over the lifetime of the chat	GPT may not be accessing data from previous sessions correctly

Pre-test 12/18/2024 (English v4)	Learners experimented with different topics & for some; the GPT went out using the PAU loop	Add explicit instructions on how to handle different topics in prompting, and additional instructions to handle all tasks within the PAU loop
Pre-test 12/18/2024 (English v4)	After v2 failure, GPT performed normally, with slight tweaks necessary	Ensure consistent language across all documents (IE 'known' vs 'target' vocabulary). Handling random topics in prompt design.
Pre-test 12/18/2024 (English v4)	Video speaking may be too fast, opportunities to simplify language, scene changes can be distracting.	Improve PAU cycle slide and use that as main graphic, include NAR mention, narrator speed to .8, revise script verbiage, test Japanese version, add stage header, rely on YouTube generated subtitles
Pre-test 12/18/2024 (English v4)	Fine-tuning documents from older versions & discrepancy between video, video scripts & process changes may be causing mistakes.	Revise and/or remove irrelevant trainings
Pre-test 12/18/2024 (English v4)	Revise old fine-tuning docs and edit them	Use edited versions as sample conversations and grading examples
Pre-test 12/18/2024 (English v4)	GPT will not be able to track personal vocabulary and scores across sessions. Sessions are defined by periods of inactivity, after which the GPT loses access to data	May need other tools to log and track discussions & personal vocabulary. Track incoming data from learners. Cumulative # of tasks correlates to use time (& other metrics,) scores are average of last 3 loops that gives a 'snapshot' of their latest scores.
Pre-test 12/23/2024 (Mandarin v4)	Prompts for beginners still don't consistently produce visuals	Produce DALL-E call-in action, may need to create 'beginner prompt' specific details in prompt design or flowchart
Pre-test 12/23/2024 (Mandarin v4)	GPT is still using characters.	Need to decide if this is allowed and clarify further, perhaps as a stage in the flowchart to ascertain if learners want to read characters.
Pre-test 12/5/2024 (Mandarin v5)	GPT was still unsure when to deliver Chinese characters or not	Can be in fixed fine-tuning

Pre-test 12/5/2024 (Mandarin v5)	GPT prompts are still inconsistently designed (extra verbiage, unnecessary explanations etc.)	Created 'interaction samples' of entire conversations for other versions
Pre-test 12/6/2024 (English v5)	Instructor group mentioned known video needs, inconsistent or disengaging prompts. Some access errors	Add further interaction samples, see video mods, encountered access errors not present in learner groups
Pre-test 12/6/2024 (English v5)	Reconsider the meaning of 'Relevance'	Include a reflection of effort in rubric. IE, relevant to the instructors objectives for the learner
Usability test 1/24/25 (English v5)	Video-only usability group tests in B1 & B1+ (some A2) understood only about 40-50% without subtitles, maybe 60% with subtitles	Produce Korean version, YouTube auto-subtitles
Usability test 1/25/2025 (English v5)	GPT was grading students repetition of prompted responses	NAR rubric (v3) Reinforce implicit corrections in the Pre-thinking stage. Consider Marco M. Usability test for fine-tuning
Usability test 1/26/2025 (English v5)	2nd video failed to load	Fixed with new 'action,' calls both videos simultaneously
Usability test 1/24/25 (English v5)	GPT was directly giving definitions of unknown words	Reinforced the provision of sample sentences for requested vocabulary in '4. Sample interaction' and in fine tuning interactions inside the GPT
Usability test 1/26/25 (English v5)	GPT is following previous instructions in ignoring mistakes during certain stages.	Based on learner feedback; revised instructions to give short, implicit corrections during the pre-thinking stage
Usability test 1/26/25 (English v5)	When grading multiples of each task type, how is the GPT choosing which to grade?	Was grading improved scores inside of PAU loops. Incorrectly grading students repetition of prompted feedback. Adjusted instructions in fine-tuning and revised NAR rubric to reflect gradual changes in learner baseline performance rather than short-term improvements within a single loop
Usability test 1/26/25 (English v5)	GPT was grading Rationale tasks inside of Sample tasks during the Use stage	Recognized a Rationale in a Sample task.

L. Appendix L

[\(Back\)](#)

GPT Performance Assessment Sheet

Below is the tally sheet for the GPT's performance. The designer observed the interactions between the GPT and the learner, categorized the GPTs responses, and discussed the observations after four usability testing sessions.

Date					
Tester name					
	Novel response	Meets expectations	Known errors	Hallucinations	Curious results
Definition	<i>An unexpected, desirable output</i>	<i>An expected, desirable output</i>	<i>An expected, undesirable output</i>	<i>An unexpected, undesirable output</i>	<i>Any category, undetermined output</i>
GPT Responses (Tally)					
GPT response notes					
Edit for fine tuning (Y/N)					

M. Appendix M

[\(Back\)](#)

Long Form Usability Questionnaire

Date		
Tester name		
Tester signature		
Process	<ul style="list-style-type: none"> • Data controlled within ChatGPT, YouTube and Google's terms of use • Resulting performance and opinion data is owned by the designer and secured privately in Google suite • Non-anonymous data may be shared with IIE and with University of Hawaii - Manoa instructors • Only anonymous data will ever be presented in any public format • Usability tests the TOOL, not the LEARNER • Do your best to THINK OUT LOUD, I will direct if necessary • Imagine how you would like to use this tool, and try it out 	
Question	Follow-up	Notes
IE Thanks for filling out the opinion survey, do you have anything you want to mention that might be easier to talk about rather than write?	Gather valuable information unsaid due to lang. barriers. Follow leads in response to the learner.	
IE While using the ChatGPT tool today, did you have any specific goal in mind?	Observe and test for different use cases, topic, grammar help, etc.	
IE The GPT skipped giving you a correction here, what do you think about that?	What type of correction do you think it should give here?	

N. Appendix N

[\(Back\)](#)

Opinion Survey Questions

Audience	Ques. Type	Data type	Question	Data analysis recommendations
All	Multi choice	Demo	GPT version: Free (short form) or paid (long form)	Learner opinions may vary based on by GPT's ability to adapt to learners in different test formats
All	Multi choice	Demo	Role: Instructor or learner	Cross analysis to compare data between expert vs. non expert
Learner: All	Short answer	Demo	First and last name	Demographic information necessary for analyzing opinion data against performance scores
Learner: All	Drop down choice	Demo	Primary language	Demographic information that highlights mistakes specific to a linguistic group
Learner: All	Drop down choice	Demo	Language level	Demographic information used to calibrate the performance of the GPT tool, define learner perceptions and preferences by level
Learner: All	Long answer	Demo	Tell us about you previous language learning experiences	Optional question that informs learner characteristics in prior language educational experiences and motivation
Learner: Exp.	Multi choice	Demo	Experienced (Y/N)	Sort students, compares results by prior experience to in/validate both instructor and experienced learner data
Learner Exp.	Likert	Learning	The new introduction video teaches the same important things as the old version	Primary instructional objective. Determines if learners believe missing components are necessary for learning. Highly relevant to the tool's efficacy for future learners, thus considered learning data
Learner: Exp.	Likert	Learning	Overall, this version is BETTER than the old version	Primary instructional objective. Determines overall sentiment around new methodology. Highly relevant to the tool's efficacy for future learners, thus considered learning data
Learner: Exp.	Long answer	Learning	If you DISAGREE with any of the statements above, why? What is missing? How can we improve the new version?	Primary instructional objective. Allows learner suggestions. Highly relevant to the tool's efficacy for future learners, thus considered learning data. May highlight known missing components in word webbing or dictionary use.

Learner: All	Likert	Usability	I was able to understand the introduction video	Suggestions in pacing, speed, narrator accent, imagery, etc. may arise in open ended questions. Cross analyzed against language level and learner performance
Learner: All	Likert	Usability	The introduction video clearly explained how to use the GPT tool	Suggestions in video edits related to presentation of the GPT usage may arise in open ended questions.
Learner: All	Likert	Usability	The messages from the ChatGPT teacher were good for my language level. Not too easy & not too hard.	Perceived value of GPT's prompt delivery and adjustment to learner level
Learner: All	Likert	Usability	The messages from the ChatGPT teacher gave me helpful corrections or tips on language use	Perceived value of GPT's ability to enhance language skills
Learner: All	Likert	Attitude	This ChatGPT tool will help me build my vocabulary	Perceived value of the GPT's effect on vocabulary acquisition
Learner: All	Likert	Attitude	This ChatGPT tool will improve my overall language use	Perceived value of the GPT's effect on overall language learning
Learner: All	Likert	Attitude	This ChatGPT tool was fun or interesting to use	Actual value of the GPT's ability to engage learners
Learner: All	Likert	Attitude	I would recommend this to a friend who wants to learn a language	Correlates with the actual value of scaling the GPT
Learner: All	Long answer	Usability Attitude	If you DISAGREE with any of the above statements, why? How can we improve this tool?	Allows learner suggestions
Instructor	Likert	Usability	The introduction video language and vocabulary was appropriate for the language level	In/validates verbiage in video for intended levels (B1-C2)
Instructor	Likert	Usability	Students will understand how to use this tool after watching the introduction video.	In/validates actual value of video design, formatting and content
Instructor	Likert	Usability	The ChatGPT teacher's messages and instructions are appropriate for the students language level	Actual value of the GPT's prompt delivery and adjustment to learner level

Instructor	Likert	Learning	The scores that the ChatGPT gives the student accurately reflect their skill level	In/validates actual value of GPT's assessment decisions
Instructor	Likert	Learning	This ChatGPT tool will help students build vocabulary	Actual value of GPT's effect on vocabulary acquisition
Instructor	Likert	Learning	This ChatGPT tool will improve students overall language use	Actual value of GPT's effect on language learning
Instructor	Likert	Attitude	Students will find this ChatGPT tool fun and interesting to use	Perceived value of GPT's ability to engage learners
Instructor	Likert	Attitude	I would recommend this tool to my students	Correlates with the actual value of scaling the GPT
Instructor	Long answer	All	If you DISAGREE with any of the above statements, why? How can we improve this tool?	Allows instructor suggestions

Analysis notes:

1. Some questions delineate perceived value versus actual value for comparison between instructor audience segments and learner segments, allowing for a balanced understanding of the GPT's ability to meet language learning standards and learner needs. Simply put:
 - a. Perceived value: Includes questions from the perspective of a non-expert opinion
 - b. Actual value: Includes questions from the perspective of an expert opinion.

O. Appendix O

[\(Back\)](#)

Novelty, Accuracy and Relevance (NAR) Rubric

Below is a summary of the GPT's standards for different tasks from the NAR Rubric. Details of this rubric can be found here: [NAR Rubric](#) and [NAR Rubric- Full Text](#)

	Connection	Assumption	Rationale	Sample
Novelty	Creative ideas or experiences	Creative guesses with flexibility	Unique reasoning or insights	Original sentences or new vocab situated in a novel context
Accuracy	Correct vocabulary and grammar	Correct guesses using context	Clear, grammatically correct logic	Grammatically accurate use of new vocabulary
Relevance	Stays on topic consistently	Context appropriate guesses	Logical, focused reasoning	Appropriate word use for context

P. Appendix P

[\(Back\)](#)

Sample NAR Score report

Below is a sample of the reports that the GPT sends to the learner. The GPT gives the total number of tasks completed as a metric of overall use over time. It gives the average scores in all categories for the lastest 3 PAU loops, this gives a ‘snapshot’ of most recent performance to compare over time. Summaries and other qualitative data is compiled via a ChatGPT for analysis. The report gets copied and pasted into the linked Google Forms survey:

“ChatGPT said:

Here’s your score report, champ! Make sure to copy and paste this into [this Google Form](#) for analysis.

NAR Score Report

Total Tasks Completed:

- **Connection Tasks:** 2 (Average NAR: Novelty 6, Accuracy 7, Relevance 7)
- **Assumption Tasks:** 2 (Average NAR: Novelty 7, Accuracy 6, Relevance 7)
- **Rationale Tasks:** 1 (Average NAR: Novelty 7, Accuracy 6, Relevance 6)
- **Sample Tasks:** 2 (Average NAR: Novelty 7, Accuracy 8, Relevance 8)

Summary

- **Strengths:**
 - Great sense of humor and natural conversational style—your sentence about traveling light was a hit!
 - Accurate use of vocabulary like "traveling light" and "deadline."
 - You quickly improved structure in business contexts with professional tone adjustments.
- **Areas to Improve:**
 - Watch out for small grammar slips (e.g., "must to" → "must").
 - Continue practicing formal email phrasing to ensure top-notch professionalism.
 - Keep tightening your sentence structure for clarity in business communication.
- **Recent Highlights:**
 - Successfully used "traveling light" in a creative and humorous context.
 - Delivered a polished response to a business question about deadlines.

Great work! Let me know if you want to dive deeper into any specific area or continue with more tasks.”

Q. Appendix Q

[\(Back\)](#)

Summary of Feedback and Recommendations

Below is a summary of edits the authors “Editing and Research Assistant” ChatGPT has made, as of 12/14/2024. This is the short form of a larger document that gave the GPT specific directions for future edits:

Key Edits Observed

1. **Grammar and Clarity:**
 - Most grammatical corrections and simplifications were implemented.
 - Some verbosity and redundant phrasing remain in sections, requiring more precise suggestions in future edits.
2. **Cohesiveness and Flow:**
 - Macro-level transitions between sections were improved, but subsection transitions occasionally lacked fluidity.
 - Linking ideas across subsections more explicitly would enhance readability.
3. **Clarity and Concision:**
 - Concision improved significantly, but a few areas still over explain or repeat concepts.
 - Future feedback should pinpoint repetitive phrases and split overly long sentences.
4. **APA 7 Compliance:**
 - Citations and references are well-formatted and consistent with APA 7 standards.
 - Cross-referencing within the document could be more standardized.

Recommendations for Future Editing

- A. **Specificity:** Provide detailed, actionable suggestions for exact sentences or transitions needing improvement.
- B. **User Preferences:** Adapt edits to balance formal academic tone with conversational clarity where appropriate.
- C. **Structural Feedback:** Focus on subsection transitions and overall organization to ensure logical flow.
- D. **Iterative Monitoring:** Continuously compare edits against prior feedback to refine alignment and catch overlooked opportunities.

Key Takeaways

- A high implementation rate reflects strong collaboration and alignment.
- Targeting transitions and concision in future edits will further improve the document.
- Maintaining this iterative and adaptable approach will support consistent, high-quality outcomes.”

R. Appendix R

[\(Back\)](#)

Table R1

Average satisfaction score in usability tests by category with standard deviation, minimum and maximum values.

Category	Question	Avg.	SD	Min	Max
Video satisfaction	Understandability	4.25	0.96	3.00	5.00
	Clearly explained the GPT tool	3.50	0.58	3.00	4.00
GPT output satisfaction	GPT's provided corrections	4.75	0.50	4.00	5.00
	GPT as a vocab builder	4.75	0.50	4.00	5.00
	GPT improves SLA	4.50	0.58	4.00	5.00
	GPT is engaging	4.75	0.50	4.00	5.00
	CR process (Inst. only)	5.00	N/A*	5.00	5.00
Critical Reading process satisfaction	Features	4.00	1.00	3.00	5.00
	Process improvement	3.00	1.00	2.00	4.00
	Recommendability	4.50	0.58	4.00	5.00

Note: This table demonstrates the resulting average scores participants gave the intervention on a Likert scale one to five, where 5 is highly satisfied.

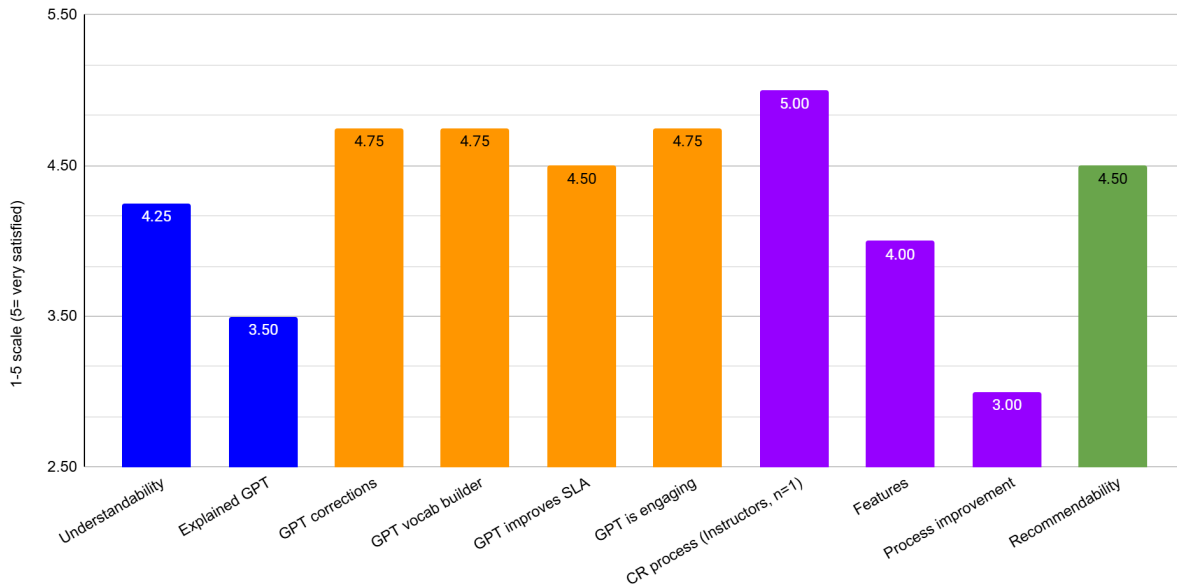
*This question only applied to instructors and the sample size was not large enough to provide SD data

Figure R1

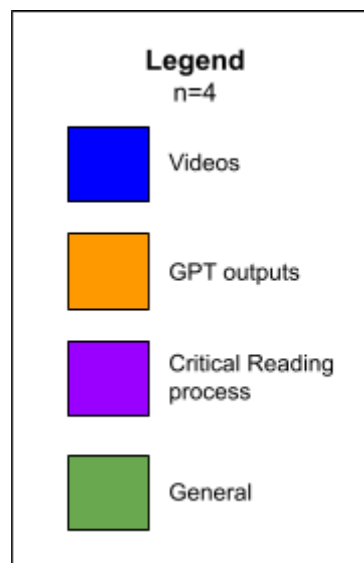
Average satisfaction score in usability tests by category

Critical Reader Usability Tests

Learner satisfaction with videos, GPT outputs and overall Critical Reading process.



Note: Critical Reading process as a whole questions was only relevant for instructors (n=1)



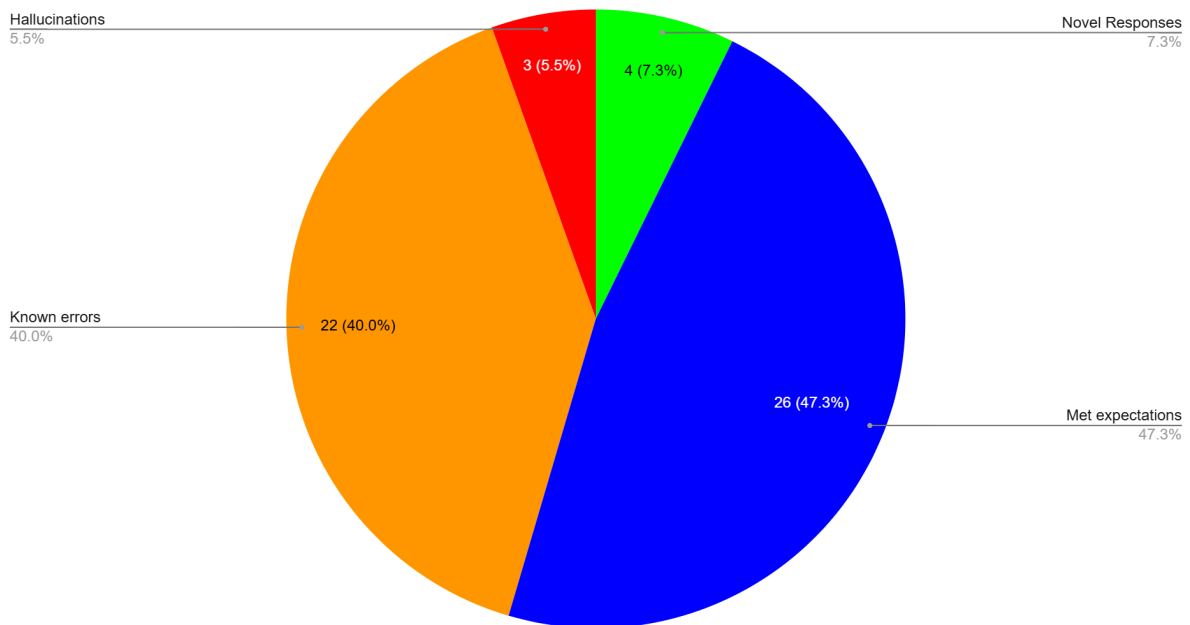
S. Appendix S[\(Back\)](#)**Table S1***Total GPT-learner interactions during four usability tests*

	Desirable	Undesirable
Expected	Met expectations (n=26)	Known errors (n=22)
Unexpected	Novel responses (n=4)	Hallucinations (n=3)

Note: GPT responses (n=55) in (n=4) usability testing sessions

Figure S1*Analysis of the GPT's responses in usability tests***Critical Reader GPT Performance Usability Tests**

Analysis of the GPT's performance during usability tests



Note: 4 test participants resulted in 55 total GPT responses

Legend

n=4

- Hallucinations
(unexpected & undesirable output)
- Known errors
(expected & undesirable output)
- Novel responses
(unexpected & desirable output)
- Met expectations
(expected & desirable output)

T. Appendix T

[Back](#)

Table T1

Total tasks completed by type from pretest (baseline) to post test (task type)

Language level	n=	Baseline Connections	Baseline Assumptions	Baseline Rationales	Baseline Samples
A2	1	2 +2	1 +1	0 +0	1 +2
B1	4	12 +3	7 +2	4 +0	9 +2
B1+	6	13 +6	9 +2	4 +0	9 +4
B2	3	9 +2	5 +1	3 +1	6 +3
C1	2	4 +2	3 +1	2 +0	3 +2
Total	16	40 55	25 32	13 14	28 41

Table T2

Total tasks completed by type pretest (baseline) to post test (task type), adjusted for sample size

Language level	n=	Baseline Connections	Baseline Assumptions	Baseline Rationales	Baseline Samples
A2	1	2.0 +2.0	1.0 +1.0	0.0 +0.0	1.0 +0.5
B1	4	3.0 +0.8	1.8 +0.5	1.0 +0.0	2.3 +2.0
B1+	6	2.2 +1.0	1.5 +0.3	0.7 +0.0	1.5 +0.7
B2	3	3.0 +0.7	1.7 +0.3	1.0 +0.3	2.0 +1.0
C1	2	2.0 +1.0	1.5 +0.5	1.0 +0.0	1.5 +1.0
Total	16	2.5 +0.9	1.6 +0.4	0.8 +0.1	1.8 +0.8

U. Appendix U

[Back](#)

Table U1

Δ in median of learner average scores by task type from pretest (baseline) to post test (task type).

Language level	n=	Baseline Connections	Baseline Assumptions	Baseline Rationales	Baseline Samples
A2	1	69 +0.2	68 +0.2	00 +0.0	74 -0.1
B1	4	75 +0.3	73 +0.1	71 +0.1	77 +0.2
B1+	6	74 +0.3	73 -0.1	70 +0.2	77 +0.1
B2	3	78 +0.2	76 +0.1	75 -0.1	81 +0.1
C1	2	80 +0.3	79 +0.1	78 -0.2	78 +0.5
Total	16	75 +0.3	73 +0.1	71 +0.0	77 +0.1

Note: Mean scores across tasks intentionally includes outliers by missed task, median here are a good measure of central tendency

Table U2

Δ in mean of learner median scores by dimension from pretest (baseline) to post test (task type).

Language level	n=	Baseline Novelty	Baseline Accuracy	Baseline Relevance
A2	1	67 +0.1	68 +0.4	71 +0.2
B1	4	69 +0.1	75 +0.3	79 +0.0
B1+	6	68 +0.1	73 +0.2	76 -0.0
B2	3	74 +0.1	78 +0.3	78 +0.2
C1	2	75 +0.2	83 +0.0	80 +0.2
Total	16	70 +0.1	75 +0.2	77 +0.1

Note: Median scores across dimensions intentionally exclude outliers by missed task, averages here are a good measure of central tendency

V. Appendix V

[Back](#)

Sample V1

B1+ Connection Score GPT Analysis

Inquiry: “B1+ student's average Connection score was 7.3 in pre-tests and 7.7 in post tests. Tell me in as much detail as possible, exactly what kind of improvements led to that increase. Use grammar terms where necessary. What specific competencies did learners demonstrate in that average increase?”

GPT response: That ~0.4 increase in the Connection score (from 7.3 to 7.7) in your B1+ students likely indicates improvements in several specific linguistic areas.

Here’s a detailed breakdown of what that increase represents in grammatical terms and cognitive processes:

1. More Complex Sentence Structures

- **Prior to the increase (Score ~7.3):**
 - Students were able to **connect ideas** using **simple** and **compound sentences** (e.g., "I like apples because they are sweet").
 - They often relied on **basic conjunctions** ("and," "but," "because") but struggled with **subordinate clauses** (e.g., "If I were a farmer, I would like to harvest bananas").
- **Post increase (Score ~7.7):**
 - Students demonstrated greater use of **complex sentences** by **embedding subordinate clauses** (e.g., "Since organic produce is healthier, I always try to buy it when possible").
 - They improved in their ability to use **relative clauses** ("The bananas, which were freshly harvested, tasted amazing") instead of simpler structures like "The bananas are fresh."

2. Improved Use of Conditional Structures

- **Prior (7.3):** Limited use of conditionals, often using **Type 1 conditionals only** ("If I study hard, I will improve my English").
- **After (7.7):**
 - More students correctly applied **Type 2 conditionals** (e.g., "If I were a farmer, I would harvest mangoes").
 - Some even attempted **Type 3 conditionals** (e.g., "If I had known about organic farming earlier, I would have studied it in school"), even if imperfect.

3. Better Use of Cohesion & Coherence

- **Before the increase:**
 - Responses had a tendency to **jump between ideas** with minimal transition markers.
 - Students often relied on **short, disjointed statements** (e.g., "I eat bananas. They are sweet. I like them.").
 - **After the increase:**
 - Improved use of **discourse markers** and transitions such as "however," "therefore," "as a result," "in contrast," etc.
 - Responses were more **fluid and logically connected** (e.g., "I eat bananas because they are sweet; however, I also enjoy other tropical fruits when they are in season.").
-

4. Expansion of Lexical Range

- **Before:**
 - Students tended to **repeat basic vocabulary** they were comfortable with (e.g., "good," "bad," "big," "small").
 - Overuse of **generic verbs** (e.g., "do," "make," "get") instead of precise vocabulary.
 - **After:**
 - Increased ability to **rephrase and expand explanations** with more sophisticated vocabulary (e.g., "nutritious" instead of "healthy," "preserve" instead of "keep").
 - Introduction of **collocations and idiomatic phrases** (e.g., "a well-balanced diet" instead of "healthy food," "a step in the right direction" instead of "a good idea").
-

5. Stronger Awareness of Contextual Meaning (Pragmatics)

- **Before:** Literal interpretations of words were more common.
 - Example: "Harvest means 'available to consume' because bananas change color."
 - **After:** Greater awareness of **metaphorical and extended meanings**.
 - Example: "Harvest means 'to collect' because farmers gather food when it's ready."
-

6. Fewer Errors in Article & Preposition Use

- **Before:** Students often omitted articles or used incorrect prepositions:
 - "Bananas is good fruit" (should be "Bananas are a good fruit").
 - "I am interested for farming" (should be "interested in farming").
- **After:** More precise usage:
 - "Bananas are a staple fruit in tropical countries."
 - "I am interested in sustainable farming methods."

7. Increased Creativity & Personalization

- **Before:** Responses were formulaic and repetitive.
 - "I like bananas because they are sweet."
 - **After:** More creative and personal connections.
 - "I choose bananas because they remind me of my childhood in Brazil, where my grandmother would make banana cake every Sunday."
-

Conclusion: What That 0.4 Increase Means

That **0.4 jump in the Connection score** suggests that learners are making **more natural, personal, and contextually appropriate connections** to new vocabulary. They are using **more varied grammar structures, applying more advanced linking strategies, and expanding their vocabulary use**. This improvement is crucial for progressing from **B1+ to B2-level proficiency** and beyond.

Sample V2

A2 and B1 Accuracy Score GPT Analysis

Inquiry: I want you to tell me in as much detail as possible, what grammatical, cognitive and linguistic processes contributed to these results: A2 level learners scored a baseline Accuracy score of 6.8 and saw a +0.4 increase in the post test while B1 students scored a baseline Accuracy score of 7.5 and saw a +0.3 increase in the post test. What exactly contributed to that increase?

GPT Response: The +0.4 increase in accuracy for A2 learners and the +0.3 increase for B1 learners in the post-test can be attributed to a combination of grammatical refinement, cognitive processing improvements, and targeted linguistic reinforcement within the PAU loop.

1. Grammatical Factors Contributing to Accuracy Gains

Since accuracy scores assess the correct use of vocabulary and grammar, the PAU loop provided direct opportunities for improvement in key areas.

A2 Level (+0.4 Accuracy Increase)

A2 learners typically struggle with word order, verb conjugation, and prepositions. The accuracy gain likely stemmed from:

- **Reduction of Grammar Errors:** Regular feedback on incorrect structures (e.g., "*He go to school*" → "*He goes to school*") helped correct common A2 mistakes.
- **Improved Subject-Verb Agreement:** Exposure to correct forms in structured feedback cycles likely helped reduce errors like "*She don't like it*" → "*She doesn't like it*".

- **Better Use of Articles & Prepositions:** Explicit correction of errors such as *“I go store”* → *“I go to the store”* reinforced proper sentence construction.
- **Increased Accuracy in Simple Past Tense:** A2 learners often struggle with irregular past tense verbs (*go* → *went*, *see* → *saw*), so targeted feedback likely helped reinforce correct usage.

B1 Level (+0.3 Accuracy Increase)

B1 learners generally make fewer grammar mistakes than A2 learners, so their accuracy gains likely resulted from refinement rather than fundamental corrections.

- **More Accurate Sentence Structures:** B1 learners may have improved their ability to structure complex sentences correctly.
 - **Better Use of Modals:** Feedback likely helped reduce errors in modal verbs (*should*, *could*, *must*), such as *“He must to go”* → *“He must go”*.
 - **Improved Word Choice Precision:** B1 learners refined their vocabulary accuracy, replacing near-synonyms (*big* → *enormous*) with more contextually appropriate words.
 - **Reduced Overgeneralization of Grammar Rules:** At B1, learners might overapply rules (*He explained me* instead of *He explained to me*), and structured corrections helped eliminate these errors.
-

2. Cognitive Learning Mechanisms Enhancing Accuracy

Both A2 and B1 learners benefited from cognitive reinforcement strategies embedded in the PAU loop.

A2 Level

- **Stronger Pattern Recognition:** Regular feedback reinforced correct structures, making them easier to recognize and recall.
- **Error Awareness Development:** Learners became more aware of their mistakes and self-corrected more often.
- **Chunking & Formulaic Language Use:** Exposure to common sentence patterns (e.g., *“I would like to...”*) helped learners internalize accurate constructions.

B1 Level

- **Improved Monitoring & Self-Correction:** B1 learners refined their ability to recognize and avoid habitual errors.
- **Higher Fluency in Producing Correct Forms:** By cycling through multiple prompts, learners internalized correct structures and required fewer corrections.
- **More Contextual Accuracy:** At this stage, learners refined how they applied grammar and vocabulary in nuanced ways.

3. PAU Loop's Role in Accuracy Gains

The PAU loop was crucial in improving accuracy because it ensured that learners repeatedly encountered correct forms through structured feedback.

1. Pre-Think Stage: Learners activated prior knowledge and formulated safe, familiar responses.
2. Assume Stage: Learners guessed at new vocabulary, and errors were corrected immediately.
3. Use Stage: Learners applied new words in context, receiving explicit feedback on errors.

This progressive correction cycle ensured that errors were systematically reduced.

4. Interaction & Engagement Impacting Accuracy

The conversational, challenge-based approach ensured that learners remained engaged, making corrections more memorable.

- **Sass and Humor in Corrections:** Playful feedback likely reduced learners' resistance to correction and increased retention.
 - **Personalized Examples:** Learners remembered corrected forms better because they were connected to their own interests and experiences.
-

5. Why A2 Learners Improved More than B1 Learners

The 0.4 increase in A2 accuracy (compared to 0.3 for B1) is expected because:

- **A2 Learners Had More Foundational Errors to Correct:** A greater number of mistakes meant greater room for improvement.
 - **B1 Learners Were Already More Accurate:** Their mistakes were fewer and often more subtle, making gains incremental rather than dramatic.
 - **Lower-Level Learners Benefit More from Direct Corrections:** At A2, explicit grammar correction has a higher impact because learners are still developing foundational accuracy.
-

Conclusion

- **A2 learners improved by +0.4 because they had more frequent grammatical errors and benefited from explicit corrections.**
- **B1 learners improved by +0.3 because their accuracy gains came from fine-tuning complex structures rather than fixing major mistakes.**
- **The PAU loop's structured correction cycle ensured repeated exposure to correct forms, leading to measurable accuracy improvements.**

This structured, feedback-driven process successfully enhanced grammatical precision and error recognition, contributing to increased accuracy scores.

W. Appendix W

[Back](#)

Table W1

Average satisfaction score in learner performance tests by category with standard deviation, minimum and maximum values.

Category	Question	Avg.	SD	Min	Max
Videos	Understandability	4.06	0.77	3.00	5.00
	Clearly explained the GPT tool	4.13	0.81	2.00	5.00
	GPT's provided corrections	4.69	0.48	4.00	5.00
GPT outputs	GPT as a vocab builder	4.31	0.70	3.00	5.00
	GPT improves SLA	4.38	0.72	3.00	5.00
	GPT is engaging	4.31	0.79	2.00	5.00
Critical reading process	CR process (Inst. only)	5.00	N/A*	5.00	5.00
	Features	4.10	0.57	3.00	5.00
	Process improvement	4.00	0.82	2.00	5.00
	Recommendability	4.38	0.72	3.00	5.00

Note: This table demonstrates the resulting average scores participants gave the intervention on a Likert scale one to five, where 5 is highly satisfied.

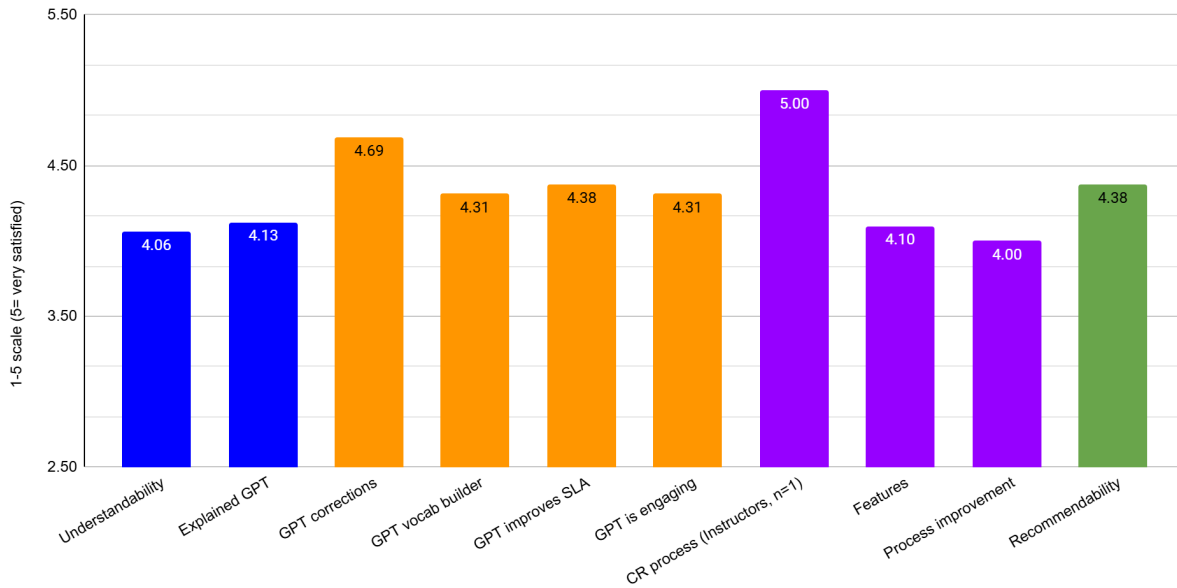
*This question only applied to instructors and the sample size was not large enough to provide SD data

Figure W1

Average satisfaction score in learner performance tests by category

Critical Reader Usability Tests

Learner satisfaction with videos, GPT outputs and overall Critical Reading process.



Note: ‘Critical Reading process as a whole’ question was only relevant for instructors (n=1)

Table W2

Average sentiment scores in select metrics by experience with Critical Reading (CR)

Experience with CR	n=	CR process features	CR process improvement	GPT as vocab builder	GPT improves SLA	GPT is engaging
Yes	10	4.1	4.0	4.2	4.3	4.2
No	6	N/A	N/A	4.5	4.5	4.5
Total	16	4.1	4.0	4.3	4.4	4.3

Note: Learners who were new to the process were not asked questions that referred to older versions of the intervention.

X. Appendix X

[Back](#)

Table X1

Primary languages of all test participants (n= 19)

Geo-linguistic group	Language	Count	Percent
East Asian	<i>Japanese</i>	8	42%
	<i>Korean</i>	2	11%
	Total	10	53%
Romance and other European	<i>German or Swiss-German</i>	5	26%
	<i>French</i>	1	5%
	<i>Portuguese</i>	1	5%
	Total	7	37%
South Asian and Middle Eastern	<i>Farsi*</i>	1	5%
	<i>Tamil</i>	1	5%
	Total	2	11%

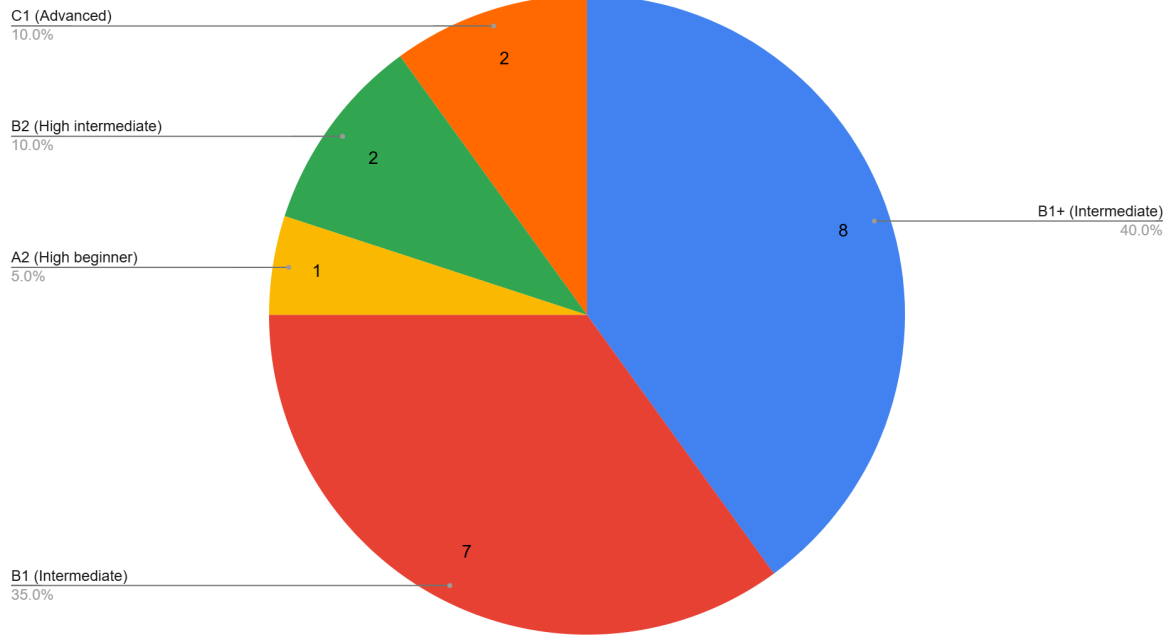
Note: One English Instructor testing as a learner was removed

*Persian

Figure X1

Language level of all test participants (n=20)

Language level of test participants



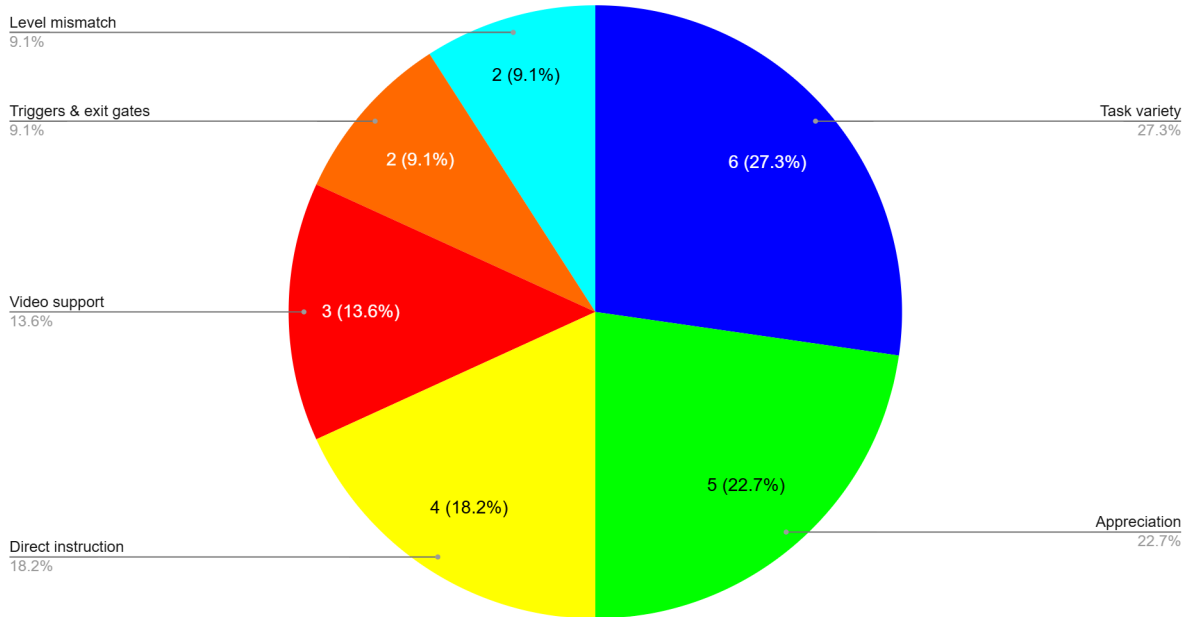
Y. Appendix Y

[Back](#)

*Learner performance test comments summarized and categorized
(n=16)*

Learner Comments

Summarized and codified comments from the Learning Performance group (n=16)



Note: 11 of 16 participants gave 22 total comments.

Legend n=16

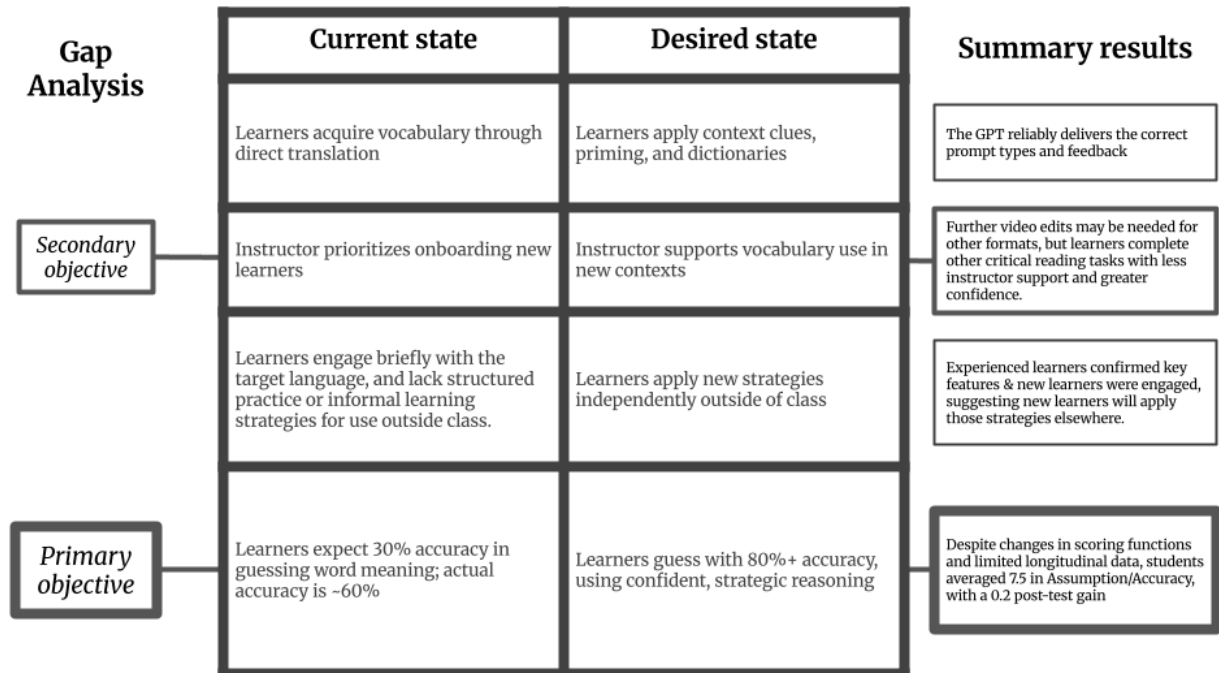
- Learners chose the wrong language level when starting
- Learners wanted more variety in instructional interfaces and tasks
- Learners appreciated the corrections and language assistance from the GPT
- Learners noted direct instruction was still preferable and necessary
- Learners mentioned that the video needs further support
- Learners would like the GPT feedback to include in-depth explanations or similar

Z. Appendix Z

[Back](#)

Figure 1

This figure summarizes the results from various usability, performance and attitudinal tests as they relate an initial gap analysis



AA. Appendix AA

Average score for Accurate Assumptions in the learner performance group (n=16)

	Avg. Assumption & Accuracy score	SD
<i>Baseline</i>	7.5	0.6
<i>Post-test</i>	7.7	0.5

GPT Use Cases, Configuration Documents, and Links

Editing and Summarizing

The author trained a ChatGPT model for editing and research assistance purposes. This model was also corrected grammar and suggested edits for clarity and brevity (see [Appendix Q](#)), and provided summaries of qualitative usability data and to provide insight into potential areas of research. This custom model also provided summaries of current work. Those summaries are included variously in this document after being edited by the author, particularly in the Literature Review section in “The Instruction; Critical Reading” section.

Intelligent Tutoring System Architecture

A rough version of the GPT used an incomplete NAR rubric. The GPT was tested numerous times with different learner levels and asked to calibrate, detail and redesign the rubric. The updated rubric was then edited by the designer and fed in the GPT for future iterations.

Learner Performance Analysis

The GPT used its built-in functions to assist in detailed analysis of student scores. See [Appendix V](#) for samples of the GPT’s analysis.

GPT Configuration Portfolio

The link here contains all the documents that configure the GPT and guide its interactions with the learner: [GPT File Drop](#) folder.

The Critical Reader

The link here brings you directly to the Critical Reader, hosted on ChatGPT: [The Critical Reader- English](#). The introduction videos can be found through interaction with the GPT.

Bibliography

- Bloom, B. S. (1984). The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring. *Educational Researcher*, 13(6), 4–16. <https://doi.org/10.3102/0013189X013006004>
- Bilingual Jobs. (2024, June 4). *Fully interactive language world map v2.0*. Retrieved February 20, 2025, from <https://bilingualjobs.io/language-world-map>
- Doroudi, S. (2023). The Intertwined Histories of Artificial Intelligence and Education. *International Journal of Artificial Intelligence in Education*, 33(4), 885–928. <https://doi.org/10.1007/s40593-022-00313-2>
- Elgort, I. (2022). Word Knowledge, Learning and Acquisition in a Second Language: Proposed Replications of Elgort (2011) and Qiao and Forster (2017). *Language Teaching*, 55(2), 260–270. <https://doi.org/10.1017/S0261444819000338>
- EnglishProfile.org. (2006). English Grammar Profile Dataset. *Cambridge University Press*.
<https://englishprofile.org/english-grammar-profile/egp-online?task=downloadXLS>
- Golonka, E. M., Bowles, A. R., Frank, V. M., Richardson, D. L., & Freynik, S. (2014). Technologies for Foreign Language Learning: A Review of Technology Types and their Effectiveness. *Computer Assisted Language Learning*, 27(1), 70–105. <https://doi.org/10.1080/09588221.2012.700315>
- Johnson, K. (2010). Expertise in Language Learning and Teaching. *ELT Journal*, 64(2), 217–218. <https://doi.org/10.1093/elt/ccp104>
- Hill, Z. (2022). MuCGEC; Chinese Grammar Error Correction. *Tianchi Datasets*.
<https://tianchi.aliyun.com/dataset/131328>
- Kapur, M. (2008). Productive Failure. *Cognition and Instruction*, 26(3), 379–424.
<https://doi.org/10.1080/07370000802212669>
- Kearsley, G., & Culatta, R. (2018a). *Andragogy (Malcolm Knowles)*. Instructional Design. <https://www.instructionaldesign.org/theories/andragogy/>

- Kearsley, G., & Culatta, R. (2018b). *Cognitive Flexibility Theory* (Spiro, Feltovitch & Coulson). Instructional Design.
<https://www.instructionaldesign.org/theories/cognitive-flexibility/>
- Khan Academy. (n.d.). *Using context clues to figure out new words* [Video]. Khan Academy. <https://www.youtube.com/watch?v=CiNggzdWkIo>
- Kim, W.-Y., & Chang, Y.-O. (2009). Linking Metacognition to Second Language (L2) Acquisition. *인문연구*, 56, 27–48. <https://www.dbpia.co.kr>
- Koedinger, K. R., Corbett, A. T., & Perfetti, C. (2012). The Knowledge-Learning-Instruction Framework: Bridging the Science-Practice Chasm to Enhance Robust Student Learning. *Cognitive Science*, 36(5), 757–798.
<https://doi.org/10.1111/j.1551-6709.2012.01245.x>
- Li, S. (2010). The Effectiveness of Corrective Feedback in SLA: A Meta-Analysis. *Language Learning*, 60(2), 309–365.
<https://doi.org/10.1111/j.1467-9922.2010.00561.x>
- Lyster, R., & Mori, H. (2006). Interactional Feedback and Instructional Counterbalance. *Studies in Second Language Acquisition*, 28(02).
<https://doi.org/10.1017/S0272263106060128>
- MacLellan, C. J., & Koedinger, K. R. (2022). Domain-General Tutor Authoring with Apprentice Learner Models. *International Journal of Artificial Intelligence in Education*, 32(1), 76–117. <https://doi.org/10.1007/s40593-020-00214-2>
- Nicholas, C., Pashler, H., Vul, E., Wixted, J., & Rohrer, D. (2006). Distributed Practice in Verbal Recall Tasks: A Review and Quantitative Synthesis. *Psychological Bulletin*, 132(3), 354–380. <https://escholarship.org/uc/item/3rr6q10c>
- Nickow, A., Oreopoulos, P., & Quan, V. (2020). *The Impressive Effects of Tutoring on PreK-12 Learning: A Systematic Review and Meta-Analysis of the Experimental Evidence* (No. w27476; p. w27476). National Bureau of Economic Research.
<https://doi.org/10.3386/w27476>

- Nwana, Hyacinth S. (1990). Intelligent Tutoring Systems: An Overview. *Artificial Intelligence Review*, 4(4). <https://doi.org/10.1007/BF00168958>
- OpenAI. (2023). ChatGPT (40 version) [Editorial and Research Assistant Chatbot-Created by Sean Lydon Minton].
<https://chatgpt.com/share/67299051-c898-800e-86c0-369cfa34030a>
- Oxford, R. L., Rubin, J., Chamot, A. U., Schramm, K., Lavine, R., Gunning, P., & Nel, C. (2014). The Learning Strategy Prism: Perspectives of Learning Strategy Experts. *System*, 43, 30–49. <https://doi.org/10.1016/j.system.2014.02.004>
- Rubin, J. (2005). The Expert Language Learner: A Review of Good Language Learner Studies and Learner Strategies. In K. Johnson (Ed.), *Expertise in Second Language Learning and Teaching* (pp. 37–63). Palgrave Macmillan UK.
https://doi.org/10.1057/9780230523470_3
- Shintani, N., & Aubrey, S. (2016). The Effectiveness of Synchronous and Asynchronous Written Corrective Feedback on Grammatical Accuracy in a Computer-Mediated Environment. *The Modern Language Journal*, 100(1), 296–319. <https://doi.org/10.1111/modl.12317>
- Shintani, N., & Ellis, R. (2013). The Comparative Effect of Direct Written Corrective Feedback and Metalinguistic Explanation on Learners' Explicit and Implicit Knowledge of the English Indefinite Article. *Journal of Second Language Writing*, 22(3), 286–306. <https://doi.org/10.1016/j.jslw.2013.03.011>
- Spiro, R. J., Feltovich, P. J., Feltovich, P. L., Jacobson, M. J., & Coulson, R. L. (1991). Cognitive Flexibility, Constructivism, and Hypertext: Random Access Instruction for Advanced Knowledge Acquisition in Ill-Structured Domains. *Educational Technology*, 31(5), 24–33. <https://www.jstor.org/stable/44427517>
- Steenbergen-Hu, S., & Cooper, H. (2014). A Meta-Analysis of the Effectiveness of Intelligent Tutoring Systems on College Students' Academic Learning. *Journal of Educational Psychology*, 106(2), 331–347. <https://doi.org/10.1037/a0034752>

- Sweller, J. (1988). Cognitive Load During Problem Solving: Effects on Learning. *Cognitive Science*, 12(2), 257–285. https://doi.org/10.1207/s15516709cog1202_4
- Tlili, A., Salha, S., Wang, H., & Huang, R. (2023). Intelligent Tutoring Systems Examined in Social Experiments—Is the Magic Gone? A Meta-Analysis. 2023 *IEEE International Conference on Advanced Learning Technologies (ICALT)*, 50–54. <https://doi.org/10.1109/ICALT58122.2023.00020>
- Wood, D., Bruner, J. S., & Ross, G. (1976). “The Role of Tutoring in Problem Solving”. *Journal of Child Psychology and Psychiatry*, 17(2), 89–100. <https://doi.org/10.1111/j.1469-7610.1976.tb00381.x>
- Xu, Z., Banerjee, M., Ramirez, G., Zhu, G., & Wijekumar, K. (Kay). (2019). The Effectiveness of Educational Technology Applications on Adult English Language Learners’ Writing Quality: a Meta-Analysis. *Computer Assisted Language Learning*, 32(1–2), 132–162. <https://doi.org/10.1080/09588221.2018.1501069>