

# Creating and Crafting CLIL Content Using AI Tools:

## A New Frontier<sup>1</sup>

Paul Wadden, Istvan Zoltan Zardai, Sophie Otsuru, and Hiroaki Umehara  
*Faculty of International Liberal Arts, Juntendo University*

---

This article offers one answer to a key question posed by the organizers of the Fifth International Symposium on Academic Writing and Critical Thinking: How can teachers and students make optimal use of AI? It recounts the authors' steps and struggles at a major Japanese university to use Large Language Models—in this case, ChatGPT—to create level-appropriate, content-based reading and listening passages for a semi-intensive English program. The authors explain how they first identified gaps in content materials in their liberal arts CLIL curriculum, and then underwent the process of crafting prompts that would generate multiple versions of the same passage to meet the needs of students at different ability levels (TOEFL ITP scores ranging from 350 to 550). The authors note how suitable CLIL content in fields such as Economics, Sociology, Natural History, Earth Science, and Health and Medicine are seldom available from commercial publishers, and they explore the promise that AI-generated content has for creating optimal learning materials. They also illustrate the limitations and the pitfalls of AI-generated materials: namely, the occasional failure of AI to follow directions, the rare but real tendency for hallucination of answers, the potential for plagiarism, and the need to carefully check AI-generated questions and answers for errors. The authors observe that the challenges that their teaching staff and students face are similar to those at many other universities across Japan, and they conclude that ChatGPT can be extremely useful when focused on appropriate topics using artfully crafted prompts. AI tools have great promise for generating level-appropriate texts in the right style and register that can then be further adapted for optimal reading and listening materials.

---

*Keywords:* AI, CLIL, Content-Based Curriculum, Japanese higher education

---

<sup>1</sup> Based upon the presentation “Creating and Modulating Course Content Using AI Tools” Delivered at the 5th Symposium on Academic Writing and Critical Thinking (AWCT) Nagoya University February 16, 2024

*Artificial intelligence is not a substitute for human intelligence; it is a tool to amplify human creativity and ingenuity.* – Fei-Fei Li, Stanford Institute for Human-Centered Artificial Intelligence (Pogla, 2024)

## **Introduction**

*How can teachers and students make optimal use of AI?* This paper provides one answer to this key question—posed by the organizers of the Fifth International Symposium on Academic Writing and Critical Thinking—as it explores the tentative, strategic use of Large Language Models (LLMs) for content creation. It particularly focuses on deploying ChatGPT for generating level-appropriate, content-based reading and listening passages—and accompanying comprehension and discussion questions—for a semi-intensive English program at a major university in Tokyo. The challenges that our teaching staff and students face in the Faculty of International Liberal Arts are similar to those at many other universities across Japan, and the explorations we are engaging in, some of the pitfalls we are encountering, and the potential solutions we are finding may be useful to other programs and faculty.

## **Curricular Backdrop**

During their initial year of university study, the 240 first-year students in our faculty take a 6-koma (6 x 90 minutes) weekly set of English courses. Using the results of a TOEFL ITP placement test, they are assigned to classes in roughly three levels—intermediate-advanced, intermediate, and lower-intermediate, corresponding roughly with CEFR B2, B1, and A2 levels. The courses are organized in a unified curriculum, meaning that all groups use shared content materials, study the same six liberal arts content modules in sequence, and take the same content tests monthly at the end of each completed learning unit. At the close of the academic year, the TOEFL ITP is again administered as a post-test to measure improvement in academic English skills; it is also administered a third time at the end of students' second year. (Wadden, 2021) Among the chief goals of students' English study in the Faculty of International Liberal Arts are to achieve a level of proficiency at which they can take the *gakubu's* newly developed EMI courses and productively study abroad in one of the university's overseas programs.

The wider national and institutional context of our English program is that 56 out of 805 universities in Japan are promoting liberal arts (Daigaku Times, 2020) and “more than 40 percent of Japanese universities now offer specialist content courses [EMI] taught in English.” (Brown 2023, p. 3) MEXT is extensively promoting EMI initiatives through the Global 30 Project, the Go Global Japan Project, and the Top Global University Project (Brown & Bradford, 2019) as well as through Kakenhi grants like the 4-year extensive grant we received (21K00714; Wadden 2021-2025) to create a new curriculum.

One of the premises of our newly established CLIL curriculum is the supposition that academic content grounded in the traditional liberal arts is fundamental to students' gaining greater academic proficiency and corollary higher TOEFL scores. As the executive director of the TOEFL at ETS (the company which makes and markets the TOEFL) observes, “Most items that you will encounter on a TOEFL test tend to be drawn directly from university-level textbooks, from the courses that students would typically encounter in a first- or second-year liberal arts class” (Gopal as quoted in Moody, 2020). At present, the ETS website (2024) states

that the reading test is “designed to measure the ability to read and understand short passages similar in topic and style to those read in courses taught at North American universities and colleges.” Prior to curriculum reform, our first-year English curriculum—like those at many universities in Japan—relied principally upon monolingual, multi-skill EFL textbooks and their learning activities in speaking, listening, reading, writing, and vocabulary. From the beginning to the end of their first-year, students gained an average of 14 TOEFL points. However, after creating our own reading and listening materials in six subject areas—Health and Medicine, Zoology, Human Evolution, Earth Science, Economics, and Sociology—the English program significantly improved students’ TOEFL scores in comparison to the previous 7-year average. In 2022, under the newly implemented curriculum, students’ scores rose 140% over the previous average; in 2023 the score gain was 108% higher than the first-year students’ historical average. However, these gains came at a cost. Since very little content-based material is available commercially, the teachers and curriculum designers wrote what was equivalent to two large coursebooks of study materials—the six liberal arts modules combined run more than 400 pages in length. These content modules were initially composed at two levels of difficulty: a basic level with fewer passages and a more limited range of sub-topics, language structures, and vocabulary, and an intermediate-advanced version with more reading and listening passages that explored a wider range of sub-topics within each discipline. However, the writing burden was enormous, and the creation of varied supporting materials also demanding. As a result, after the first year the two-tiered approach was dropped as unsustainable. This led to the persistence of the problem that the two-level approach was implemented to solve<sup>2</sup>: Satisfying students’ language-learning needs by keeping them in the zone of proximal development with materials that were not too easy and not too difficult. Put differently, the range of English ability between students scoring from 350 to 550 on the TOEFL ITP is too wide for one-size-fits-all materials to bridge.

Enter ChatGPT.

### **Up and Running: Generating New Materials**

Following the end of the 2023 academic year, members of the teaching team began experimenting with ChatGPT to generate content and questions for revising and expanding its liberal arts modules. The teachers reasoned, from their previous experience, that the labor-intensive process of drafting and revising (and then laying out and printing) multiple versions of liberal arts texts for different levels of students was beyond their capacity. But they decided to experiment in some portions of the module by including two versions of the same reading passage one after another (along with similarly modulated comprehension and discussion questions). They reasoned that advanced classes could start with the more challenging passage and skip the less challenging one, or they could use it for extensive- or reinforcement-reading. The lower-intermediate students, on the other hand, could begin with the less challenging passage and then ratchet up to the higher level one after first grasping the fundamental concepts and key vocabulary. Many of our lower-intermediate students’ vocabulary can be estimated to range somewhere between 1,600-3,000 words when they enter our program, hence they are not yet able to read and focus primarily on content (Hirsh and Nation, 1992). In fact, some students

---

<sup>2</sup> At the beginning of the first year of reform, the curriculum designers even tried a 3-level approach, which they gave up on after a month.

may encounter new vocabulary in almost every line of text, and often more than one. Hence, repetition is still crucial for them.

Connected by Zoom online over break, teachers first identified gaps in the current module materials and then in small groups began using ChatGPT prompts to generate text. In the early days of coding and information management, programmers had an adage, “Garbage in, garbage out” (GIGO). Its successor still lives on in AI deployment. The quality of the input prompt largely determines the quality of the output response. As many have pointed out, the art of using LLMs hinges upon crafting prompts that elicit the answers and texts desired, and that produce them in their proper register, length, level of detail, amount of exemplification, and format. Here, for example, is a progression of prompts the authors used for content creation for their Health and Medicine module as they began their efforts. They wanted a short introductory passage on the differences between infectious diseases and lifestyle diseases.

- In two paragraphs explain the difference between infectious diseases and lifestyle diseases at the 5th grade level (with subsequent prompts requesting the 6th and 7th grade levels).
- In two paragraphs explain the difference between infectious diseases and lifestyle diseases at the CEFR B1 level for non-medical students.
- In three paragraphs, explain the differences between infectious diseases and lifestyle diseases at the CEFR B1 level for non-medical students. Write several general discussion questions about the text as well as 5 multiple-choice comprehension questions.
- Using simple language and the 2,800 words of the NGSL, in 3 paragraphs at the CEFR B1 level for non-medical students, explain the differences between infectious communicable diseases and chronic lifestyle diseases and provide examples of each kind. Then, write 3 discussion questions followed by 5 TOEFL style multiple-choice reading comprehension questions with 4 answers each. One of the questions should be a vocabulary question.

There were more steps in the evolution of the prompts than those listed above but you get the picture. By the time we reached the last iteration, we were able to generate some fairly good text and strong questions. But this process was enabled by teachers collaborating together to identify key terms for the prompt that would elicit higher-quality generated text—for example, some teachers from North America opted to modulate the text by grade level and some from Europe felt CEFR was a better way to tune the language band. In the end, as teachers refined and expanded the prompt, ChatGPT generated content materials on the differences between infectious and lifestyle diseases that are extremely hard to find in commercial textbooks, and it adapted the passages and the questions to two different language proficiency levels (B1 and A2 in this case).

### **HAL, are you okay? Why doesn't ChatGPT follow simple directions?**

The process was hardly flawless. Like the servant Caliban in Shakespeare's “The Tempest,” ChatGPT isn't great at following directions and sometimes seems to even have a stubborn mind of its own. For example, when generating text for the Economics module, we instructed it:

*In two paragraphs explain the difference between inflation and deflation at the CEFR B1 level. Then, write 5 discussion questions followed by 5 TOEFL style*

*multiple-choice reading comprehension questions with 4 answers each. One of the questions should be a vocabulary question.*

In response, the LLM produced only a single paragraph, not two as requested, and it failed to include a vocabulary question, even though it was explicitly instructed to do so. Although the vocabulary was well-modulated in the B1 bandwidth, a few words and phrases could be flagged as closer to B2 level, such as “various factors” and “circulation.” (You can even ask ChatGPT to identify which words are in what CEFR bandwidth, although you shouldn’t completely trust its answers!) Here is the single economics paragraph it produced:

*Inflation and deflation are two important concepts in economics that describe changes in the overall level of prices in an economy. Inflation occurs when prices of goods and services rise over time, meaning that consumers need more money to purchase the same things. This can happen because of various factors such as increased demand, rising production costs, or government policies that increase the amount of money circulating in the economy. In contrast, deflation happens when prices decrease over time, so consumers need less money to buy the same goods and services. Deflation can occur due to factors like weak demand, falling production costs, or government policies that reduce the amount of money in circulation.*

The discussion questions included the following:

1. How does inflation affect consumers' purchasing power?
2. What are some potential causes of deflation in an economy?
3. How might inflation and deflation impact businesses differently?
4. Can you think of any examples of periods of inflation or deflation in recent history?
5. How do governments typically respond to high levels of inflation or deflation?

Although this passage was too brief for our purposes, the text was generated in about 10 seconds. It is well-written, artfully structured, relatively formal in tone, offers clear definitions of key terms “inflation” and “deflation,” and elaborates on those key terms with multiple explanations of their causes. Read it out loud and ask yourself how challenging it would be for you to compose a concise paragraph explaining these concepts in such a measured tone. And how long would it take you? While the initial ChatGPT passage was too short and lacked a requested question type, refinements to the prompt soon led to a longer more suitable multi-paragraph reading passage with the requested vocabulary and reading comprehension questions.

Another shortcoming to keep an eye on in ChatGPT-generated learning materials is comprehension questions which read well but upon scrutiny have no right answer. Using the process outlined above, we generated two passages on “Blue Zones”—places where people are unusually healthy and live long lives—at the CEFR A2 level (left) and B1 level (right).

**Version 1 (CEFR A2)**

1. Blue zones are special places where people live longer and healthier lives than in other parts of the world. One blue zone is

**Version 2 (CEFR B1)**

1. Blue zones are regions around the world where people are known to live longer, healthier lives compared to the global

Okinawa, Japan. People there eat lots of vegetables, tofu, and fish, which they believe helps them stay healthy. They also like to do activities like gardening and walking every day. In Okinawa, having strong connections with friends and family is very important, and this makes people feel happy and supported.

2. Another blue zone is Ikaria, which is in Greece. People in Ikaria eat a Mediterranean diet, which means they enjoy foods like fresh vegetables, fruits, and olive oil. They also like to relax and take naps in the afternoon. In Ikaria, people have close relationships with each other and spend a lot of time talking and laughing together. This helps them feel less stressed and more connected.

3. In Loma Linda, California, there's another blue zone, but it's a bit different. Here, many people are Seventh-day Adventists, which means they follow certain Christian religious beliefs. They eat a plant-based diet, which includes foods like whole grains, beans, nuts, and fruits. People in Loma Linda also like to exercise and do things to relax, like spending time with friends and going to church. Their strong Christian faith and sense of community help them feel happy and healthy. Even though these blue zones are in different parts of the world and have different ways of living, they all share the idea that eating well, staying active, and having close relationships are important for a long and healthy life.

average. One such zone is Okinawa, Japan, where residents have a diet rich in vegetables, tofu, and fish, which contributes to their longevity. In addition to their diet, Okinawans maintain an active lifestyle through daily activities like gardening and walking. Social connections are also crucial in Okinawa, with strong community bonds and support networks playing a significant role in the well-being of its residents.

2. Another blue zone is Ikaria, Greece, known for its Mediterranean diet and relaxed way of life. Ikarians consume a variety of fresh vegetables, fruits, and olive oil, which are believed to contribute to their longevity. Unlike Okinawa, Ikarians emphasize leisurely activities and stress reduction, with afternoon naps and socializing being common practices. The strong sense of community in Ikaria fosters close relationships and a sense of belonging, promoting overall health and happiness among its inhabitants.

3. In contrast, Loma Linda, California, is a blue zone with a unique cultural aspect: the Seventh-day Adventist community. Unlike the other blue zones, religion plays a significant role in the community's well-being, with Christian practices such as Sabbath observance and communal worship contributing to a sense of purpose and inner peace among its members. In addition, residents follow a plant-based diet, emphasizing whole grains, legumes, nuts, and fruits. Regular physical activity and stress management techniques are also parts of their lifestyle.

Despite their geographical and cultural differences, these blue zones share commonalities in promoting healthy diets, active lifestyles, and strong social

	connections as keys to longevity and well-being.
--	--

For the CEFR B1 passage, ChatGPT generated the following two multiple-choice comprehension questions.

1. Which of the following is NOT mentioned in the passage as a characteristic of the Okinawan diet?
  - A. The eating of vegetables, tofu, and fish
  - B. Consumption of processed foods
  - C. Using olive oil in food preparation
  - D. Emphasizing the importance of whole grains
  
2. What sets apart Ikaria, Greece from Okinawa, Japan in lifestyle practices?
  - A. Emphasis on social connections
  - B. Relaxed way of life
  - C. Stress reduction techniques
  - D. Religious practices

At first glance, these seem like solid questions. However, for the first question, the key word “NOT” needs to be deleted for the question to have a correct answer, otherwise three of the answers are arguably correct. For the second question, there is no correct answer. An “Emphasis on social connections” (A) and “Relaxed way of life” (B) are true of both Ikaria and Okinawa. Stress reduction techniques (C), if broadly interpreted, is also true of both Blue Zones. Finally, “Religious practices” (D) is irrelevant. Of course, comprehension questions can be crafted from these two drafts (in the first, “NOT just needs to be deleted), but it is important to scrupulously check any ChatGPT materials that have discrete answers.

More serious errors and even hallucinations also sometimes appear. According to other ChatGPT investigators, one only has to use the names of well-known people and add a theory, and ChatGPT comes up with nonexistent theories. For example, philosophers Timothy O’Connor and Hong Yu Wong authored a classic paper in the philosophy of mind hypothesizing the mind as an emergent property of physical systems (O’Connor and Wong, 2005). When one prompts ChatGPT to discuss the [nonexistent] “O’Connor-Wong view of emergent electricity” the LLM obligingly produces the following fascinating but nonsensical hallucination:

*The O’Connor-Wong view of emergent electricity is a theory proposed by physicists John J. O’Connor and S.C. Wong. This theory suggests that electricity may be an emergent phenomenon, meaning it arises from the collective behavior of many interacting particles rather than being fundamental at the smallest scales of matter.*

*In traditional physics, electricity is described by fundamental forces and particles, such as electrons and protons, interacting through electromagnetic fields. However, O’Connor and Wong proposed that at certain scales or under specific conditions, such as in certain condensed matter systems or complex materials, electricity may emerge as a macroscopic property from the collective behavior of electrons and other charged particles.*

*This view challenges the conventional understanding of electricity as solely a fundamental force and suggests that emergent phenomena could play a role in its behavior in certain contexts. The O'Connor-Wong view has sparked interest and debate in the physics community and continues to be explored in various research areas, including condensed matter physics and materials science.*

This AI explanation sounds informed and convincing. The problem is that there is no O'Connor-Wong view of electricity at all and something that doesn't exist certainly has not sparked debate in the physics—or philosophy—communities. One thing to note is that ChatGPT used concepts found in the actual work of O'Connor and Wong and mixed these with key vocabulary of theories of electricity. We experienced a similar result when referees for a journal asked us to clarify the difference between CBI (Content-Based Instruction) and CLIL for a paper on our newly created curriculum. We prompted ChatGPT 3.5 to explain the difference between CBI and CLIL approaches and requested that it provide quotations and citations. The LLM hallucinated quotations on the difference between these approaches, and fabricated two citations using contemporary scholars' names to support what it claimed were distinctions between them.

Another example of somewhat dodgy text emerged when we generated materials for our Earth Science module using the prompt, "Describe different types of volcanoes at CEFR B1 level" (see left) and at the "CEFR B2 level" (see right).

<b>Version 1 (CEFR B1)</b>	<b>Version 2 (CEFR B2)</b>
<p>Volcanoes come in different shapes and sizes, and they can be classified into several types based on their shape, eruption style, and the type of magma they produce. Here are some common types of volcanoes:</p> <ol style="list-style-type: none"><li>1. Shield Volcanoes: These volcanoes have gentle slopes and are broad in shape, resembling a warrior's shield lying on the ground. They are formed by lava flows that are not very explosive. Shield volcanoes often erupt with relatively quiet lava flows, which can travel long distances from the vent. Mauna Loa in Hawaii is an example of a shield volcano.</li><li>2. Stratovolcanoes (Composite Volcanoes): These are tall, steep-sided volcanoes with a cone-shaped appearance. They are formed by alternating layers of lava flows, volcanic ash, and other volcanic debris. Stratovolcanoes can have explosive eruptions, producing ash</li></ol>	<p>Certainly! Volcanoes are fascinating natural phenomena, and they come in various types, each with its unique characteristics. Here are descriptions of different types of volcanoes:</p> <ol style="list-style-type: none"><li>1. Shield Volcanoes: These volcanoes have gentle slopes and wide bases, resembling a warrior's shield lying on the ground. They are formed by the accumulation of fluid lava flows that spread out over large areas. Shield volcanoes typically have non-explosive eruptions, with lava flowing steadily from vents. Examples include Mauna Loa and Mauna Kea in Hawaii.</li><li>2. Stratovolcanoes (Composite Volcanoes): Stratovolcanoes are tall, conical mountains characterized by steep sides and a central vent or crater. They are formed by alternating layers of lava flows, volcanic ash, and other volcanic debris. Stratovolcanoes can produce</li></ol>



clouds, pyroclastic flows, and lava flows. Mount Fuji in Japan and Mount St. Helens in the United States are examples of stratovolcanoes.

3. Cinder Cone Volcanoes: Cinder cone volcanoes are small, steep-sided volcanoes formed by the accumulation of volcanic debris ejected during eruptions. The ejected material consists of volcanic ash, cinders, and volcanic bombs, which pile up around the vent to form a cone-shaped structure. Cinder cone eruptions are generally short-lived and relatively small in scale. Paricutin in Mexico is an example of a cinder cone volcano.

4. Calderas: Calderas are large, basin-shaped depressions that form when a volcano collapses into its emptied magma chamber during a massive eruption. These eruptions can be extremely explosive and can eject large volumes of volcanic material. Calderas can vary in size, with some spanning tens of kilometers across. Yellowstone Caldera in the United States is one of the most famous examples of a caldera.

These are just a few examples of the different types of volcanoes found around the world. Each type has its own distinct characteristics, eruption styles, and associated hazards. Understanding the type of volcano is crucial for assessing the potential risks and impacts of volcanic activity in a particular area.

both explosive and effusive eruptions, often resulting in pyroclastic flows, ash clouds, and lava flows. Mount Vesuvius in Italy and Mount Fuji in Japan are well-known examples.

3. Cinder Cone Volcanoes: These are small, steep-sided volcanoes with a bowl-shaped crater at the summit. They are formed by the accumulation of volcanic fragments such as cinders, ash, and volcanic bombs ejected during explosive eruptions. Cinder cone volcanoes typically have short-lived eruptions characterized by ash clouds and lava fountains. Paricutin in Mexico is a famous example of a cinder cone volcano.

4. Calderas: Calderas are large, circular depressions that form when the magma chamber beneath a volcano empties during a massive eruption, causing the collapse of the volcano's summit. They can range in size from a few kilometers to tens of kilometers in diameter. Caldera-forming eruptions are among the most explosive and catastrophic events in nature, often resulting in widespread devastation. Yellowstone Caldera in the United States and Santorini Caldera in Greece are notable examples of calderas.

Understanding the different types of volcanoes is essential for assessing their behavior, eruption styles, and associated hazards. By studying these features, scientists can better predict and mitigate the impacts of volcanic activity on surrounding communities and ecosystems.

Introductions and conclusions resulting from each prompt, as you can see above, differ greatly. For some reason, ChatGPT performs better when outputting B1 level texts. The first sentence of the B2 level text fails to follow academic conventions closely and includes unnecessary mood setting elements (“Certainly!”) and the quasi-personal comment of volcanoes being ‘fascinating.’ Such stylistic elements have a function in teaching, blogging, personal writing, and other forms of knowledge dissemination, yet the academic texts we prepare for students generally lack such elements. In addition, AI seems to have made only

minor sentence-structure and example changes to the paragraphs describing the 4 types of volcanoes. The changes it did make were surprisingly minor; for example, the last line of the description of stratovolcanoes reads in the B1 level text (paragraph 2), “Mount Fuji in Japan and Mount St. Helens in the United States are examples of stratovolcanoes,” while the last line on stratovolcanoes in the B2 version changes the order and modifies one of the two examples to read “Mount Vesuvius in Italy and Mount Fuji in Japan are well-known examples.”

Even more to our surprise, the output of ChatGPT on volcanoes seemed to have uncannily similar structure, wording, and examples as our already existing Earth Science reading on volcanoes. This might have been one instance where we encountered some of the more morally and legally potentially problematic aspects of LLMs: intellectual property theft and plagiarism. For recent developments on regulating these aspects of LLMs see for example European Parliament (2023), and for a more detailed discussion of IP infringements by LLMs see Eurojust (2023). In fact, ChatGPT 3.5 and some other LLMs are unable to name the sources they draw on, which makes the origin of the information unidentifiable. We ourselves tested this by prompting ChatGPT to name the sources it worked with, resulting in a message claiming that this was not possible. We only noticed the potential plagiarism due to one of our members being the author of our passage on the same topic—a text that had previously appeared in a widely used TOEFL preparation text. Interestingly, a plagiarism scan report on “Plagiarism Detector.net” flagged the ChatGPT-generated volcano content as 20% plagiarized compared to 5% for the original passage already in our module.

Yet another issue to consider is whether to use ChatGPT 3.5, the free version, or ChatGPT 4.0, the paid version, which costs \$20 a month to subscribe to. When one begins to use 3.5 for larger amounts of text, it soon throttles back the output as an incentive to get the user to upgrade to 4.0. For example, when generating English definitions and example sentences for content vocabulary (inflation, deflation, fiscal, recession etc.), 3.5 soon begins to reduce the number of words, definitions, and example sentences that it will generate at one time from a prompt. In our case, it at first processed the requested materials for 40-50 words at a time, but that soon slowed to 20-25, and then further to 15-20. For our work asking ChatGPT to produce initial drafts of passages three to four paragraphs in length, however, we found that the free version 3.5 worked adequately. Yet another concern for some topics is how up-to-date the information is, which is quite important for contemporary political, economic, and social issues. After considering the challenge of deciding which AI engine to use, we decided to go to the horse’s mouth and ask ChatGPT itself about the difference between the LLM knowledge-base of the two versions.

*What is the difference between chatgpt 3.5 and 4.0?*

ChatGPT 4.0 Answer	ChatGPT 3.5 Answer
<p>The primary differences between GPT-3.5 and GPT-4 revolve around improvements in performance, understanding, and generation capabilities. Here's a detailed breakdown of the main differences:</p> <ol style="list-style-type: none"> <li>1. Scale and Training Data: GPT-4 is</li> </ol>	<p>As of my last update in January 2022, I don't have specific details on the differences between GPT-3.5 and GPT-4, as GPT-4 was not released by that time. However, typically with new iterations, advancements are made in areas such as:</p> <ol style="list-style-type: none"> <li>1. Model Architecture: There could be</li> </ol>

trained on a much larger dataset compared to GPT-3.5, and incorporates more diverse and recent data. This includes a broader range of internet text up to its knowledge cutoff in January 2022. GPT-3.5, by contrast, was trained with data only up until 2021.

2. **Model Size and Architecture:** GPT-4 is generally larger in terms of the number of parameters compared to GPT-3.5. This increase in scale typically contributes to improved understanding and generation capabilities.
3. **Performance on Complex Tasks:** GPT-4 tends to perform better on more complex tasks involving reasoning, problem-solving, and understanding nuanced text. It is more adept at handling detailed and multifaceted prompts.
4. **Handling of Subtle Contexts:** GPT-4 shows improved performance in understanding subtle nuances in language and can maintain context over longer conversations more effectively than GPT-3.5.
5. **Reduced Biases and Errors:** GPT-4 has been designed to reduce biases and is better at avoiding the generation of harmful or misleading content. It also makes fewer factual errors compared to earlier models.
6. **Multimodal Abilities:** GPT-4 has capabilities in both text and image modalities, allowing it to generate and understand not just textual content but also to work with images in certain configurations (like DALL-E).

These enhancements make GPT-4 more reliable and versatile for a wider range of applications, from conversation and content creation to more specialized tasks like coding assistance and data analysis.

changes or improvements in the underlying architecture of the model, potentially allowing for better understanding and generation of text.

2. **Training Data:** Updates might include using larger or more diverse datasets for training, which could lead to better performance and comprehension across various topics and languages.
3. **Performance:** GPT-4 may demonstrate improvements in generating coherent, contextually relevant responses, reducing errors, and handling more complex queries.
4. **Capabilities:** New versions might introduce additional capabilities, such as better handling of nuanced language, improved context retention over longer conversations, or enhanced ability to generate creative or informative responses.
5. **Fine-tuning and Specialization:** There could be advancements in techniques for fine-tuning the model for specific tasks or domains, improving its suitability for a wider range of applications.
6. **Ethical and Safety Considerations:** OpenAI might introduce updates aimed at mitigating potential biases, promoting responsible use, and ensuring the model's outputs are ethical and safe.

Keep in mind that these are general expectations based on the evolution of AI models, and specific details would depend on the actual changes implemented by OpenAI in GPT-4, which I cannot provide as of my last update.

The general knowledge which provides the foundation for our liberal arts content topics (economics, earth science, zoology, etc.) has changed little since 2021 or since January 2022: bacteria and viruses still have the same differences; inflation and deflation the same causes and inverse relationship to each other; the principle types of volcanoes haven't changed. However, using ChatGPT 3.5 for researching subjects for our own professional interests, such as CLIL versus CBI, or EMI at Japanese universities, or recent scholarship on generative AI (such as best use cases or relevant regulations) will omit more recently published scholarship. Alert: even 4.0 has more than a two-year lag!

Despite some of these shortcomings, our main findings are that ChatGPT is extremely useful when focused on appropriate topics using well crafted prompts. It can generate level-appropriate texts in the right style and register that can then be adapted for reading and listening passages. It isn't consistent across levels, and all texts need to be checked and edited, both for factual accuracy and textual quality. Nevertheless, using the LLM can save immense time and effort, modulate texts for students at different levels, and save students considerable money on the purchase of textbooks, if the university, like ours, is willing to foot the bill for self-printed texts. It should be noted that when generating text it is prudent to try multiple levels (such as CEFR B1 and B2); we even found that for listening passages, prompting the LLM to generate text at A2 level output resulted in better content.

Finally, the question remains, what is the best AI product to use for generating content and language-learning materials most suited to our students' levels. When we began, we intended to also experiment with LLMs such as Google's Gemini (previously Bard), Perplexity AI, and Quillbot to compare their performance to Open AI's ChatGPT. However, our time constraints and the time pressure of generating, refining, and readying text for our own students by the beginning of the academic year meant that we had to postpone comparison of AI tools for a future day, and to the future efforts of other colleagues.

### **Questions, Conclusion, Closing Thoughts**

There are a host of questions that we have left unanswered. For example, how useful are LLMs for evaluating student essays? What AI programs can best create lecture and listening audio for students? How can LLMs be used to create more personalized and autonomous language-learning programs for students? Is it feasible and conducive to academic honesty to teach students themselves how to use AI to generate texts for their English classes? In general, we found that ChatGPT was extraordinarily fast and relatively proficient at producing strong drafts of learning materials adapted to our students' language levels. Revising those drafts, refining language-learning activities, and checking answers still required significant time and effort. But compared to finding acceptable commercial texts (a nearly impossible task) and to writing the materials from scratch (an enormous undertaking), ChatGPT definitely delivered. As questions regarding the intellectual property status of the texts are resolved, better know-how for crafting prompts is developed, and a greater understanding of how LLMs affect evaluation and coursework are available, ChatGPT and its competitors might emerge as some of the best tools for generating level-appropriate text for CLIL- and CBI-based curricula.

## References

- Brown, H. (2023). "Recruiting for Japanese EMI programs: Insights from online job posts." In P. Ferguson, B. Lacy, & R. Derrah (Eds.), *Learning from Students, Educating Teachers—Research and Practice*. JALT.
- Brown, H. & Bradford, A. (2019). "Teaching subject content through English: CLIL and EMI courses in the Japanese university." In P. Wadden and C. Hale *Teaching English at Japanese Universities: A New Handbook* 101-108.
- Daigaku Times. (2020, May). "リベラルアーツを取り入れている大学一覧." Daigaku Times. Accessed 23/11/2023 [https://times.sanpou-s.net/special/vol36\\_1/5.html](https://times.sanpou-s.net/special/vol36_1/5.html)
- ETS. (2024 May). ETS TOEFL ITP: Reading Comprehension Sample Questions. Accessed 13/05/2024 [https://toefl-samples.ets-rschtech-prod.c.ets.org/toefl\\_www/toefl\\_itp/test\\_preparation/sample\\_questions/level1\\_section3\\_reading\\_comprehension.html](https://toefl-samples.ets-rschtech-prod.c.ets.org/toefl_www/toefl_itp/test_preparation/sample_questions/level1_section3_reading_comprehension.html)
- Eurojust. (November 2023). "Generative Artificial Intelligence. The Impact on Intellectual Property Crimes." <https://www.eurojust.europa.eu/sites/default/files/assets/generative-ai-impact-to-ip-crimes.pdf>
- European Parliament. (2023). "EU AI Act: First Regulation on Artificial Intelligence." European Parliament. Accessed 11/05/2024. <https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence>
- Moody, J. (2020). TOEFL tips for prospective international students. U.S. News and World Report. Retrieved September 23, 2020 <https://www.usnews.com/education/best-colleges/articles/toefl-tips-for-prospective-international-students>
- O'Connor, T. and Wong, H. Y. (2005). "The Metaphysics of Emergence." *Noûs* 39 (4), 658–678. doi: 10.1111/j.0029-4624.2005.00543.x
- Pogla, M. (5 February 2024). "80 Artificial Intelligence Quotes Shaping the Future." AutoGPT. Accessed 12/05/2024 <https://autogpt.net/most-significant-famous-artificial-intelligence-quotes/>
- Wadden, P. (2021). "Liberal arts, first-year English, and TOEFL: An illustrative case for curricular reform at Japanese universities." *Juntendo University Journal of Global Studies*, 6, 127-136.
- Wadden, P. (2021-2025). Kakenhi. "Addressing Japan's International TOEFL Ranking with an 'English for Liberal Arts' Curricular Model for Colleges and Universities." (21K00714)

## Acknowledgments

This material has been presented at the 5th International Symposium on Academic Writing and Critical Thinking on Friday February 16, 2024, at Nagoya University. We thank the Japanese Ministry of Education for its multi-year Grant-in-Aid for Scientific Research (*kakenhi*) for the project entitled "Addressing Japan's International TOEFL Ranking with an 'English for Liberal Arts' Curricular Model for Colleges and Universities" (21K00714).

## Author Bios

**Paul Wadden**, PhD in English (with a focus on writing and rhetoric), is a special project professor in the Faculty of International Liberal Arts at Juntendo University, Tokyo, where he co-coordinates a first-year English program for 240 students. He is the editor of *A Handbook for Teaching English at Japanese Colleges and Universities* (Oxford, 1993) and *Teaching English at Japanese Universities: A New Handbook* (Routledge, 2019). His articles have appeared in *TESOL Quarterly*, *ELT Journal*, *RELC Journal*, *Composition Studies*, and many other publications. His current interests include developing CLIL curricula in the liberal arts and intensive vocabulary study providing a foundation for academic language learning.

**Istvan Zoltan Zardai**, Ph.D. in Philosophy, is a lecturer in the Faculty of International Liberal Arts at Juntendo University, and a visiting lecturer at Keio University, teaching English, philosophy, and politics. He has published on the nature, explanation, and moral psychology of action; on issues regarding liability and responsibility for the behaviour of AI; and on the relationship of philosophy and politics. He is especially interested in applied ethics and political philosophy, and currently exploring and examining student engagement, the regulation of emerging technologies, and the ethics of armed conflicts and violent resistance to unjust systems.

**Sophie Otsuru** earned an MSc in Psychology of Language from Edinburgh University. Her research interests include cognitive linguistics, the use of prepositions by native and non-native speakers of a language as well as Second Language Acquisition in general. She currently teaches in the Faculty of International Liberal Arts at Juntendo University. Prior to coming to Japan, she taught English for Academic Purposes at King's College London and other British universities.

**Hiroaki Umehara** earned an M.A. in Applied Linguistics from Texas Tech University. He currently teaches in the Faculty of International Liberal Arts at Juntendo University. His research interests focus on developing grammar and vocabulary teaching materials for EFL learners. Additionally, he is the founder of a non-profit organization dedicated to community development. His work in community design emphasizes integration through play, offering canoeing activities and water-safety workshops to individuals and schools in the Kanto area.