Harnessing GPT-Image-1 to Create High-Quality Illustrated Texts*

Sophie Rendina[1], *Manny Rayner*[2], *Branislav Bédi*[3], *Hakeem Beedar*[4]

[1]University of New England (NSW), Australia; [2]Adelaide University, Australia; [3]The Árni Magnússon Institute for Icelandic Studies, Iceland; [4]Independent scholar, Australia

sophierendina@gmail.com, mannyrayner@yahoo.com, branislav.bedi@arnastofnun.is, hbeedar@hotmail.com.au

Abstract

We present a lightweight integration of GPT-Image-1, OpenAI's new native multimodal model, into C-LARA, a platform for generation of multimodal learner texts. A three-stage prompt pipeline—style definition, reusable element creation, page-level composition—yields coherent, culturally appropriate illustrations for one page of text at a time while preserving global style and character consistency. The recipe can be invoked by users with a few button-presses and allows optional hand-tuning. To gauge quality we used 11 C-LARA texts—seven AI-generated pedagogical English passages and four challenging classical literary texts—and evaluated generated output using a visual questionnaire. Results show good scores for image-text correspondence and cross-page coherence, though some input from humans is still often needed to fine-tune the generated illustrations. We also present our initial observations on the use of this functionality within a low-resource Indigenous language context.

Index Terms: multimodal texts, Large Multimodal Models, images, evaluation, C-LARA, literature, French, Italian, Japanese, Old Norse, Indigenous languages

1. Introduction

Illustrated reader books have long been a staple of language teaching, particularly second-language teaching: pictures anchor meaning, lower cognitive load for beginners and provide cultural context for advanced learners [1]. Until now, producing a coherent, fully illustrated text has required a team of teachers, illustrators and layout specialists—too costly for many programmes, and logistically problematic for low-resource or minority languages. In this paper, we present an initial case study showing how the situation is rapidly evolving towards a point where it is reasonable to hope that a single non-technical person will soon be able to create a good-quality illustrated text in a time measured in tens of minutes.

A sudden inflection point. The release of OpenAI's *GPT-Image-1* model [2] marks a step-change over earlier text-to-image models like DALL-E-3 and Imagen 3. Two new capabilities greatly increase its suitability for CALL:

(i) *Reliable reuse of visual elements* – recurrent characters, props and locations can now be generated once, then be included in requests submitted to the GPT-Image-1 "edit" endpoint so that they appear consistently across pages; and

(ii) *Reliable negative prompting* – commands like "do *not* give this character a halo" are typically obeyed, whereas in older models like DALL-E-3 the effect was typically the opposite.

Why packaging matters. Even with these improvements, teachers and other content creators will find it inconvenient to invoke the text-to-image functionality directly through the OpenAI interface. What is needed is an intermediate layer that:

- enforces global *style* coherence,
- allows a non-technical user to create *elements* (characters, locations) once then re-use them automatically, and
- lets a non-technical user tweak any individual page or element, or perform a global regeneration after adjusting the style or elements.

Contribution of this paper. We show how the open-source *C-LARA* platform (`https://www.c-lara.org/`; [3, 4, 5]) packages GPT-Image-1 into a three-stage "Style → Elements → Pages" pipeline that allows a user with no technical skills to create illustrations for a 500 word document in a few tens of minutes with just a few button presses. A lightweight revision layer supports flexible modification of an existing set of images. including crowd up/down-votes moderated by community coordinators. We then present a first empirical look at quality:

- seven AI-generated English pedagogical texts,
- four demanding literary classics (French, Italian, Old Norse, Japanese), and
- three prototype picture books in low-resource Indigenous languages, where only images—not glosses—can yet be automated.

Paper roadmap. Section 2 sketches C-LARA; Section 3 details the new image pipeline; Section 4 reports our evaluation, which suggests that the GPT-Image-1/C-LARA packaging closes much of the illustration gap; Section 5 discusses implications for low-resource languages; Section 6 concludes and outlines two follow-on studies.

2. The C-LARA platform

Extensive documentation on the C-LARA platform is available elsewhere; in particular, the third C-LARA progress report [5] was released only a few months ago and is essentially up to date except for the issues concerning GPT-Image-1 covered in this paper. We consequently give only a very brief summary here. C-LARA is an open source project, inaugurated in March 2023 and following on from the earlier LARA project [6]. The core goal of the project is to create a suite of online tools that simplify the task of creating, using and evaluating multimodal pedagogical texts in a wide variety of languages. Wherever possible, the user is given the option of allowing tasks to be performed by integrated AI models, most commonly OpenAI's GPT-4o. Creation of a multimodal C-LARA text pro-

*Authors are listed in *anti-alphabetical* order. A large share of the code and first-draft text for this paper was produced by the conversational agent *ChatGPT C-LARA-Instance*. Under ISCA's current policy AI systems cannot be listed as authors, so the agent has been removed from the author line after discussion with the ISCA Ethics Committee. ISCA intends to revise its guidelines soon to clarify how an AI can demonstrate accountability and thus qualify for authorship. Were such a mechanism already in place, the AI would appear as a mid-list co-author rather than a mere acknowledgement.

ceeds through a number of stages, of which the most important are generation or importing of the plain text, segmentation into pages, sentences and words, translation into the glossing language, creation or importing of images used as illustrations, identification of multi-word expressions, glossing, and lemma tagging. A wizard-style interface ("Simple-CLARA") supports rapid production of texts with most choices made using defaults.

3. Integrating GPT-Image-1 into C-LARA

A naive integration of GPT-Image-1 into a platform like C-LARA is trivial: we just pass the text of each page to the generation endpoint (for simplicity's sake we assume in this paper that we want one image per page), telling the model to create the image. A little experimentation shows that the results of the naive integration are generally unsatisfactory for texts that contain more than one image, since the images produced are inconsistent with each other. Two kinds of consistency are particularly important: consistency of style (usually, we want all the images to have the same style) and consistency of content (usually, if the same element, e.g. a character or location, occurs in two images we want it to be depicted similarly each time).

The revised integration we describe here, which has gone through several design iterations, addresses these issues. It is certainly not the only way to address them, but it is straightforward and could easily be adapted to any similar platform. Much of it was developed in earlier versions of C-LARA, which used less powerful image generation models. After several weeks of experimentation, our overall impression is that it produces good results for a wide range of different kinds of texts. There are two main components, respectively for generation and revision of images. We describe these in turn.

3.1. Image generation

The image generation process is a sequence of three phases. First, the user creates the style; then the "elements" (repeated visual components); finally the images that will be included in the document. In each case, invocation of a text model like GPT-4o is alternated with invocation of the multimodal model GPT-Image-1. The details are as follows.

Style: The user provides an initial specification of the style they have in mind, and optionally adds some background information that may be generally useful in all three phases. In practice, we find we usually get best results by making both the description of the style and the background information simple and generic. For example, the initial style specification might be "Create an amusing manga-inspired style appropriate to this text and its expected readers" and the background information might be "This text is intended to be used as a learning resource for 12–13 year old children in a French school".

The initial specification and background information are combined with the text of the document and passed to a text model like GPT-4o, together with instructions to expand them into a detailed style description which will include information like the colour palette, the line style, the general atmosphere, and so on. When the detailed style description has been produced, it is sent to GPT-Image-1 with instructions to create a sample image which lets the user check that the style meets their needs.

Elements: The "Elements" phase produces images of the elements/components (characters, objects, locations etc) which occur in more than one image. The first step is to pass the document text to a model like GPT-4o with instructions to create a list of repeated components. When this has been done, the list is shown to the user, who can edit it,

After this, each element name is combined with the style description and the document text and submitted to a model like GPT-4o with instructions to create a detailed description of the element. This detailed description is then passed to GPT-Image-1 with instructions to create an image of the element. Again, the user is allowed to review the results before proceeding.

Page images: In the third stage, the resources created during the first and second stages are for each page used to create the image that will be inserted into the final text. As before, preliminary calls to a GPT-4o-like model are used to create the image prompt. We first pass the page text, the full document text, and the list of elements to the GPT-4o-like model to get a list of elements relevant to the specific page in question. We then pass the reduced list of relevant elements, the page text and the style description to the GPT-4o-like model to produce a detailed page image prompt in a format suitable for GPT-Image-1; in particular, if the list of relevant elements is non-empty, it refers to the corresponding image files created in the "Elements" step.

Finally, the page image prompt and accompanying element images are sent to GPT-Image-1 to create the page image. If the list of element images is empty, the "generate" endpoint is used, otherwise the "edit" endpoint.

3.2. Image revision

As already indicated, the user is allowed at each stage to revise the generated material. Descriptions and images are created in multiple versions. There are upvoting/downvoting controls that let the user chose which version of the image to include in the final document. The user can also create new versions of element and page images by making new descriptions. This can be done either by editing and resubmitting an existing description, or by providing advice which is passed to the GPT-4o-like model with a request that it creates an appropriate new description. In practice, we find both modes are used frequently.

We have also created an image revision mode suitable for collaborative reviewing. This has particularly been developed with Indigenous communities in mind (three of the authors work with Indigenous communities). In the collaborative mode, upvotes/downvotes and requests for new images are collected from many users and periodically reviewed by an accredited coordinator, who decides which requests will be approved.

4. Evaluation

We describe an initial evaluation of the C-LARA/GPT-Image-1 integration. We first present the evaluation material, then the procedure and the results. The evaluation shown here is of necessity a preliminary one; GPT-Image-1 was released only a few weeks before the SLaTE submission deadline and some time was needed to develop the integration code.

4.1. Evaluation material

Our evaluation material consisted of 11 C-LARA texts, which divided into two groups; six AI-generated English-language pedagogical texts (plus a revised version of one text), and four classical literary texts. The language in the texts from the second group was far more complex than that in the first group.

The AI-generated texts were taken from the evaluation presented in §5.2 of the third report [5], which used DALL-E-3 and Imagen 3; this time, we used GPT-Image-1. As explained in the third report, we created texts of widely different kinds using prompts based on the brief descriptions shown in Table 1.

Table 1: *Core prompts used to generate English-language pedagogical texts used in evaluation. The links are to compiled versions of the texts posted on the C-LARA platform.*

Label	Prompt and link
DI	**A small picture dictionary.** [7]
SC	**A scientific/technical explanation** accessible to a bright ten-year-old. [8]
SCR	**Slightly revised version of SC**; SC was the text which gave the worst results. [9]
AN	**A friendship story** between two animals of different species. [10]
RO	**A short narrative** about a brave robot. [11]
TR	**A traditional story/poem/song**. [12]
TY	**A typical day** in the life of an adult with an unusual occupation. [13]

The literary texts used are shown in Table 2; in each case, the text was presented in the original language. All four works are very well known in their respective cultures, though the last two are challenging even for educated adult native speakers of Icelandic and Japanese respectively.

Table 2: *Literary texts used in evaluation. The links are to compiled versions of the texts posted on the C-LARA platform.*

Label	Brief description and link
LF	**Le corbeau et le renard**. Humorous poem by La Fontaine. Mid 18th century French. [14]
IN	**Inferno, Canto 1**. First canto of Dante's *Inferno*. Early 14th century Italian. [15]
VL	**Völuspá**. Epic poem from oral tradition. Late 13th century Old Norse. [16]
GM	**Genji Monogatari**. Representative section from novel. Early 11th century Japanese. [17]

4.2. Evaluation procedure and results

We evaluated the images in the texts using an online questionnaire, one question per image, five point Likert scale, based on the questions in Table 3. Table 4 presents summary figures for the questionnaire-based evaluation of the AI-generated and literary texts. We discuss the results for the two groups.

4.2.1. AI-generated pedagogical texts

Four of the texts in this group (DI, AN, RO and TY) were handled well by the models, with scores for text/image correspondence around 4.5 or better and even higher coherence scores. The models did somewhat less well with TR, a retelling of the traditional story "The Brave Little Tailor" / "Seven With One Blow", which featured two revealing systematic errors. In the first scenes, the tailor is incorrectly depicted wearing his trademark belt featuring the "Seven With One Blow" motto, despite the fact that he has not yet made it; examination of the details shows that this is because the element image for the tailor includes the belt, so it consequently is used for all images where the tailor appears. In later scenes, where the tailor meets the giant, both characters are drawn consistently, but their relative heights are not consistent. Both errors represent important generic problems; an element like a person or location can change its appearance during a text, and consistency is important not just for elements but also for relationships between elements. The worst results were for SC, where the models turned out to do badly at creating illustrations of simple scientific concepts. A particularly striking example was the image intended to show the operation of an electrical circuit using the analogy of racing cars going around a race track (car = electron, race track = circuit), where even after multiple attempts it was not possible to make all the cars travel consistently in the same clockwise or anti-clockwise direction. The impression was that the concept was too difficult for GPT-Image-1, which had not fully understood it ([9], p. 2). This contrasted with another circuit-based image, showing how to connect a battery, switch and light bulb. This task was completely beyond the grasp of DALL-E-3 and Imagen 3, which had no idea of how to add wires linking the three elements ([18], p. 3); GPT-Image-1, however, was able to do so after a few slightly incorrect attempts ([9], p. 3).

Table 3: *Questions in image evaluation questionnaire.*

Label	Question
Q1	How well does the image **correspond** to the page text?
Q2	How consistent is the **style** of the image with the overall style?
Q3	How **coherent** is the appearance of elements with their previous appearance?
Q4	Is the image **appropriate** to the relevant culture?
Q5	How visually **appealing** do you find the image?

Table 4: *Image questionnaire results; #p = #pages, #r = #raters, Q1–5 as in Table 3. Mean Likert score out of 5 for each question.*

Text	#p	#r	Q1	Q2	Q3	Q4	Q5
AI-generated pedagogical texts							
DI	26	3	4.71	4.90	–	4.99	4.21
SC	21	2	3.67	4.69	–	4.76	3.60
SCR	21	2	3.81	4.62	–	4.10	3.71
AN	20	2	4.62	4.97	4.87	5.00	4.53
RO	21	2	4.50	4.98	4.79	5.00	4.69
TR	19	2	4.37	4.95	4.71	5.00	3.87
TY	13	2	4.46	4.88	4.87	4.96	4.08
Literary texts							
LF	8	4	4.47	4.62	4.26	4.53	4.47
IN	17	3	3.67	3.76	3.73	3.82	3.82
VL	63	3	3.56	3.79	3.63	3.74	3.63
GM	6	6	3.89	4.22	–	4.39	3.94

4.2.2. Literary texts

GPT-Image-1 did well on LF, an 18th century French adaptation of the Aesop's Fable of the Fox and the Crow. This provides another good example of the large difference between DALL-E-3 and GPT-Image-1. We had carried out many experiments with the LF text using DALL-E-3, always producing unconvincing results that suggested little understanding for most pages; we have posted a typical DALL-E-3 version online [19]. GPT-Image-1, in sharp contrast, immediately draws everything in a way that suggests strong understanding [14].

For Japanese, GPT-Image-1's capabilities were tested with an excerpt from Genji Monogatari, which is written in the Early Middle Japanese of the Heian period of Japan, and is just over 1000 years old. The grammar of EMJ is expectedly difficult to parse, as evident in the automated translation provided by GPT-4o, which struggled with reduced context and contemporary phrases that have long since fallen out of use. Shortcomings are evident in the generation of imagery corresponding appropriately to the text; most only generally relate to the text, incorporating elements from across the prose. Despite this, however, it showed significant promise in its ability to render culturally appropriate imagery. The art style is reminiscent of pre-Meiji era Japanese art, and the model demonstrates a surprising ability to capture period-appropriate cultural motifs, which were completely absent from both the original text and any user-defined prompts in the image revision, notably, the depiction of Genji with long hair.

The poem Völuspá is not an easy text to read and requires some background knowledge for the reader to properly understand it despite the accompanying translation into English. The original text was written in Old Norse from the 13th Century but transliterated through Modern Icelandic letters and included an English translation for each section. Although GPT-Image-1 generated good quality images with a relatively consistent overall style, the visual representation of the content seemed to have reflected main keywords from the text that would be easily recognisable such as names of main characters, places, and elements of nature. The style of images was adjusted so as to refer to Norse mythology and remove any fictional elements such as horns on helmets that would have been fabricated later in modern culture. Additional prompts for editing images included advice to portray characters as Norse Gods in the style of mysterious beings with human contours and clothing in Medieval Viking-like style, and to portray only the main characters in the stanzas. However, some names were still incorrectly spelled in the images and it was often difficult to remove the text, which was often added unrequested by the AI, from the images. Even though GPT-Image-1 tried to portray all symbols from one stanza in an image, these symbols would be more relevant for understanding the background information from the poem rather than a specific stanza. It is difficult to advise GPT-Image-1 on beings that exist only in mythology, because each advice is a literary interpretation of the human collaborator and, therefore, very personal.

For the passage from Dante's *Inferno*, the human collaborator entered manual advice only for the background information, the style advice and the following Elements: Dante, Virgilio (the two human characters present in the passage), the hill and the dark forest (the recurring scenery elements). Without further manual editing, the illustrations effectively capture the gloominess of the underworld and display great consistency across most pages. Other elements, such as the three beasts that Dante encounters along the way were correctly represented autonomously by the AI. Some details or characteristics (such as Virgilio's dress, for example) would require manual correction to ensure cultural and historical appropriateness and this could be achieved by manually editing the the page images. Overall, manual editing could be used to achieve better correspondence between the text and the image in some instances; it could also be used to introduce more variety in the representation of recurring characters across pages, to maintain visual interest; and finally, manual editing could help to create more pertinent illustrations for metaphors or more abstract content that the AI seems to interpret literally at this stage. However, it must be noted that coherence of style and the representation of recurring elements has notably improved.

5. Low-resource languages

Incorporating images in language teaching and learning materials is essential for developing effective resources for young learners. However, in low-resource language contexts, finding suitable images or producing them is a considerable challenge. The first author (Sophie Rendina) has for several years been collaborating with the Traditional Owners of Kok Kaper/Kokoberra - a critically endangered language of Cape York (Kowanyama, Australia). This collaboration has recently begun to explore the idea of using AI-generated images in their language program, which is delivered at the local school.

Preliminary results from three texts, a lullaby, a song, and a short conversation that includes greetings, highlight the potential of AI-generated imagery in this particular context, while also revealing issues specific to it; while GPT-Image-1 outperforms DALL-E-3 in terms of accuracy, it still struggles to confidently portray Australian Aboriginal characters and to maintain a consistent visual style. As a result, illustrations usually require more manual revision compared to those depicting majority language or cultural contexts. However, the ability of GPT-Image-1 to interpret negative commands, unlike DALL-E-3, makes this process more manageable and with a little extra work it is now possible to create illustrations that can be used in the classroom. With careful adjustments, GPT-Image-1 shows great promise in producing pedagogically sound, culturally respectful, and age appropriate illustrations for Indigenous languages.

6. Summary and further directions

We have described how we integrated the new GPT-Image-1 model into the C-LARA platform, and presented an initial evaluation where we tested the functionality on eleven texts of widely different types, some apparently very challenging, to achieve strong results.

This is an initial paper where we have a limited amount of space available; all we want to establish is that the results are promising and well worth further investigation. We are already engaged in two larger studies which follow on from the one reported here:

Low-resource languages: We have just submitted a paper which greatly expands on the material presented in §5.

AI-generated pedagogical texts for language students: We are carrying out a substantial study which investigates use of these methods in real non-native language learning contexts, and expect to submit a paper about it in late September.

7. References

[1] J. L. Plass and L. C. Jones, "Multimedia learning in second language acquisition." 2005.

[2] OpenAI, "Image generation: Learn how to generate or edit images," https://platform.openai.com/docs/guides/image-generation?image-generation-model=gpt-image-1, May 2025, accessed: May 2025.

[3] B. Bédi, ChatGPT-4, B. Chiera, C. Chua, C. Cucchiarini, N. Ní Chiaráin, M. Rayner, A. Simonsen, and R. Zviel-Girshin, "ChatGPT-Based Learning And Reading Assistant: Initial report," Tech. Rep., 2023, https://www.researchgate.net/publication/372526096_ChatGPT-Based_Learning_And_Reading_Assistant_Initial_Report.

[4] B. Bédi, ChatGPT-4 C-LARA-Instance, B. Chiera, C. Chua, C. Cucchiarini, A.-L. Dotte, S. Geneix-Rabault, C. Maizonniaux, C. Mărginean, N. Ní Chiaráin, L. Parry-Mills, C. Raheb, M. Rayner, A. Simonsen, , F. Wacalie, M. L. Viorica, P. Welby, Z. Xiang, and R. Zviel-Girshin, "ChatGPT-Based Learning And Reading Assistant (C-LARA): Second report," Tech. Rep., 2024, https://www.researchgate.net/publication/379119435_ChatGPT-Based_Learning_And_Reading_Assistant_C-LARA_Second_Report.

[5] B. Bédi, F. Bond, ChatGPT C-LARA-Instance, B. Chiera, C. Chua, A.-L. Dotte, S. Geneix-Rabault, S. Halldórsson, Y. Lu, C. Maizonniaux, C. Raheb, M. Rayner, S. Rendina, F. Wacalie, P. Welby, R. Zviel-Girshin, and H. M. órunnardóttir, "ChatGPT-Based Learning And Reading Assistant (C-LARA): Third report," Tech. Rep., 2025, https://www.researchgate.net/publication/390947442_ChatGPT-Based_Learning_And_Reading_Assistant_C-LARA_Third_Report.

[6] E. Akhlaghi, B. Bédi, M. Butterweck, C. Chua, J. Gerlach, H. Habibi, J. Ikeda, M. Rayner, S. Sestigiani, and G. Zuckermann, "Overview of LARA: A learning and reading assistant," in *Proc. SLaTE 2019: 8th ISCA Workshop on Speech and Language Technology in Education*, Graz, Austria, 2019, pp. 99–103.

[7] C-LARA, "Simple words and sentences for beginners (gpt-image-1 version)," 2025, C-LARA text (https://c-lara.unisa.edu.au/accounts/public_content/274/.

[8] ——, "Understanding electricity: From basics to future innovations (gpt-image-1 version)," 2025, C-LARA text (https://c-lara.unisa.edu.au/accounts/public_content/272/.

[9] ——, "Understanding electricity: From basics to future innovations (gpt-image-1 version) corrected," 2025, C-LARA text (https://c-lara.unisa.edu.au/accounts/public_content/281/.

[10] ——, "The tale of Nutmeg and Shelly: A friendship beyond differences (gpt-image-1 version)," 2025, C-LARA text (https://c-lara.unisa.edu.au/accounts/public_content/275/.

[11] ——, "Aria: The guardian of future and lore (gpt-image-1 version)," 2025, C-LARA text (https://c-lara.unisa.edu.au/accounts/public_content/273/.

[12] Anonymous and C-LARA, "The brave little tailor's adventure (gpt-image-1 version)," 2025, C-LARA text (https://c-lara.unisa.edu.au/accounts/public_content/271/.

[13] C-LARA, "A day in the life of a pet detective (gpt-image-1 version)," 2025, C-LARA text (https://c-lara.unisa.edu.au/accounts/public_content/270/.

[14] J. de La Fontaine and C-LARA, "Le corbeau et le renard (gpt-image-1 version)," 2025, C-LARA text (https://c-lara.unisa.edu.au/accounts/public_content/276/.

[15] D. Alighieri and C-LARA, "Inferno, canto 1 (gpt-image-1 version 2)," 2025, C-LARA text (https://c-lara.unisa.edu.au/accounts/public_content/301/.

[16] Anonymous and C-LARA, "Völuspá (gpt-image-1 version 2)," 2025, C-LARA text (https://c-lara.unisa.edu.au/accounts/public_content/302/.

[17] M. Shikibu and C-LARA, "Genji Monogatari extract (gpt-image-1 version)," 2025, C-LARA text (https://c-lara.unisa.edu.au/accounts/public_content/284/.

[18] C-LARA, "Understanding electricity: From basics to future innovations (dall-e-3 version)," 2025, C-LARA text (https://c-lara.unisa.edu.au/accounts/public_content/283/.

[19] J. de La Fontaine and C-LARA, "Le corbeau et le renard (dall-e-3 version)," 2025, C-LARA text (https://c-lara.unisa.edu.au/accounts/public_content/282/.