

GenAIReading: Augmenting Human Cognition with Interactive Digital Textbooks Using Large Language Models and Image Generation Models

Ryugo Morita

Hosei University & DFKI GmbH
Tokyo, Japan
ryugo.morita.7f@stu.hosei.ac.jp

Ko Watanabe

DFKI GmbH
Kaiserslautern, Germany
ko.watanabe@dfki.de

Jinjia Zhou

Hosei University
Tokyo, Japan
jinjia.zhou.35@hosei.ac.jp

Andreas Dengel

DFKI GmbH
Kaiserslautern, Germany
andreas.dengel@dfki.de

Shoya Ishimaru

Osaka Metropolitan University
Osaka, Japan
ishimaru@omu.ac.jp

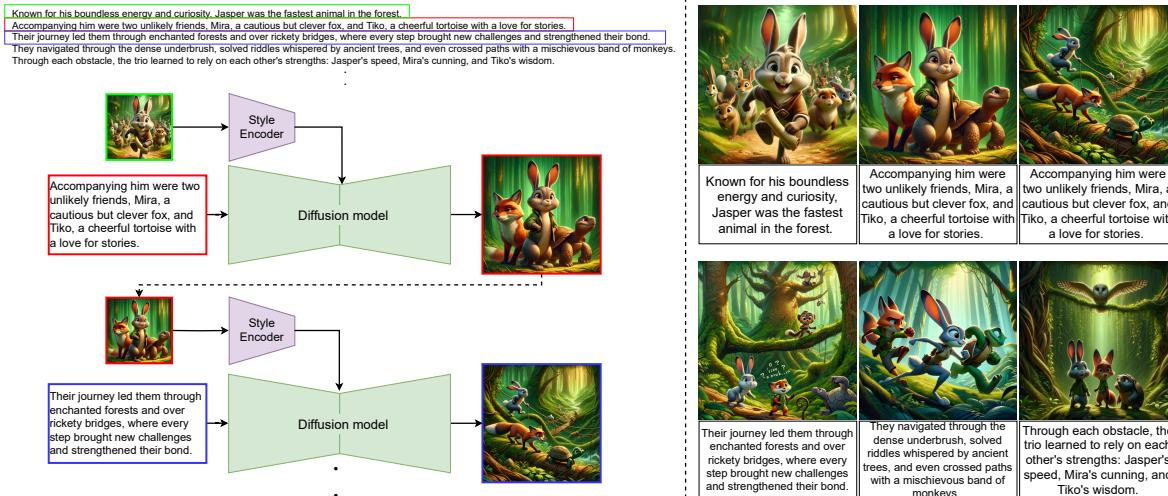


Figure 1: The model takes both a provided sentence and a preceding image as inputs to comprehend the narrative context and stylistic elements. It then generates a corresponding story image that aligns with the input sentence and the style inferred from the previous image. The images displayed on the right side demonstrate various outcomes the model produces based on different input sentences.

ABSTRACT

Cognitive augmentation is a cornerstone in advancing education, particularly through personalized learning. However, personalizing extensive textual materials, such as narratives and academic textbooks, remains challenging due to their heavy use, which can hinder learner engagement and understanding. Building on cognitive theories like Dual Coding Theory—which posits that combining textual

and visual information enhances comprehension and memory—this study explores the potential of Generative AI (GenAI) to enrich educational materials. We utilized large language models (LLMs) to generate concise text summaries and image generation models (IGMs) to create visually aligned content from textual inputs. After recruiting 24 participants, we verified that integrating AI-generated supplementary materials significantly improved learning outcomes, increasing post-reading test scores by 7.50%. These findings underscore GenAI's transformative potential in creating adaptive learning environments that enhance cognitive augmentation.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY'

© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

CCS CONCEPTS

- Applied computing → Education;
- Human-centered computing → User interface toolkits; User studies.

KEYWORDS

large language models, generative models, eye-tracking, text-to-image, prompt engineering

ACM Reference Format:

Ryugo Morita, Ko Watanabe, Jinja Zhou, Andreas Dengel, and Shoya Ishimaru. 2018. GenAIReading: Augmenting Human Cognition with Interactive Digital Textbooks Using Large Language Models and Image Generation Models. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 13 pages. <https://doi.org/XXXXXXX.XXXXXXX>

1 INTRODUCTION

Cognitive augmentation [5, 18, 31] is a pivotal concept in advancing human cognition through technology. Schmidt [31] emphasize how interactive technologies—such as remote collaboration tools, mobile computing, and machine learning—have democratized cognitive enhancement, making it widely accessible. Similarly, Clinch and Ward [5] highlights the transformative role of emerging wearables, advanced sensing modalities, and portable neuroimaging technologies in further enhancing cognitive capabilities. Together, these innovations demonstrate how technology revolutionizes human cognition, fostering intellectual growth and performance.

In educational psychology, retaining extensive textual materials, such as academic textbooks and narratives, remains a significant challenge [15]. While these materials are essential for building foundational knowledge, their text-heavy formats can hinder engagement, particularly in the digital age, where visual integration has become vital to improving educational experiences. Technological advancements—such as cameras, videos, and digital platforms—have expanded the role of visual content in promoting effective learning strategies. Paivio's Dual Coding Theory posits that encoding information in textual and visual formats enhances memorization and comprehension [24]. Mayer's research further supports this theory, showing that text integrated with visual aids significantly improves learning outcomes [21].

However, the widespread adoption of visual aids in educational materials remains limited due to the significant resources and expertise required to design high-quality, contextually relevant visuals. Traditional printed textbooks also face physical constraints that restrict the inclusion of supplementary visual content. These limitations have led educators and publishers to prioritize text-heavy formats, which often fail to address the diverse needs of learners who benefit from multimodal content [29].

We propose leveraging Generative AI (GenAI) to bridge this gap and enrich educational materials with adaptive, personalized content for human cognitive augmentation. Figure 1 represents the overall idea of our application. Our study investigates the impact of AI-generated supplementary materials—text summaries, images, and image summaries—on learners' comprehension and retention. We created concise text summaries using Text Generation AI (TGenAI), while Image Generation AI (IGenAI) converted sentences into corresponding images. We further developed a novel *Summary Image Selector* to intelligently curate the five most relevant images, ensuring alignment with the narrative flow.

Our empirical study involved diverse participants and utilized eye-tracking data to analyze engagement with AI-generated materials. The results revealed that incorporating AI-generated text summaries, images, and image summaries improved post-reading test scores by 1.25%, 4.58%, and 7.50%, respectively. This demonstrates the potential of AI-driven materials to enhance learning outcomes significantly. Additionally, we observed correlations between post-reading test scores and learners' preferences for text or image materials. Eye-tracking data highlighted that personalized educational content tailored to learners' cognitive profiles can optimize learning experiences, providing further evidence of AI's potential to address diverse learning needs.

This research lays the groundwork for developing AI-driven educational tools that automatically generate adaptive learning materials. Our framework can significantly enhance learning outcomes by transforming how content is created and tailored. Future research should expand on these findings across diverse populations and explore real-world applications of these tools in educational settings. Our key contributions are as follows:

- C1 We present a novel framework integrating AI-generated text summaries and images into educational materials, demonstrating how Generative AI can improve comprehension and retention. The innovative *Summary Image Selector* method aligns visual content with the narrative.
- C2 Through a comprehensive empirical study supported by eye-tracking data, we evaluate the effectiveness of AI-generated materials across different learner preferences, highlighting their potential to enhance educational outcomes.
- C3 We provide design guidelines for AI-driven educational tools, emphasizing optimizing visual content to accommodate diverse cognitive styles and learning preferences.

2 RELATED WORK AND BACKGROUND

In this section, we introduce prior work on use of generative AI in human-computer interaction, and human memory and comprehension augmentation.

2.1 Generative AI in Human-Computer Interaction

Generative AI, encompassing text and image generation technologies, has significantly impacted various domains within Human-Computer Interaction (HCI). Since the advent of Transformers [36], Transformer-based Large Language Models (LLMs) have become prevalent tools in HCI applications.

LLMs have been utilized in diverse fields, including personalized marketing and targeted advertising [10, 42], creative arts such as music generation [43], and the aesthetic evaluation of poetry and storytelling [37]. In visual arts, LLMs assist in creative decision-making, like analyzing color theory in interior design [14]. In education, LLMs support the "Learning by Teaching" paradigm, where students teach AI agents to reinforce their understanding while receiving feedback [1, 20]. They also aid in study planning and organizational support [33], vocabulary learning [41], and provide writing assistance [11, 25]. While LLMs have seen extensive adoption in HCI, the rise of diffusion-based image generation models introduces new possibilities and challenges. These models have

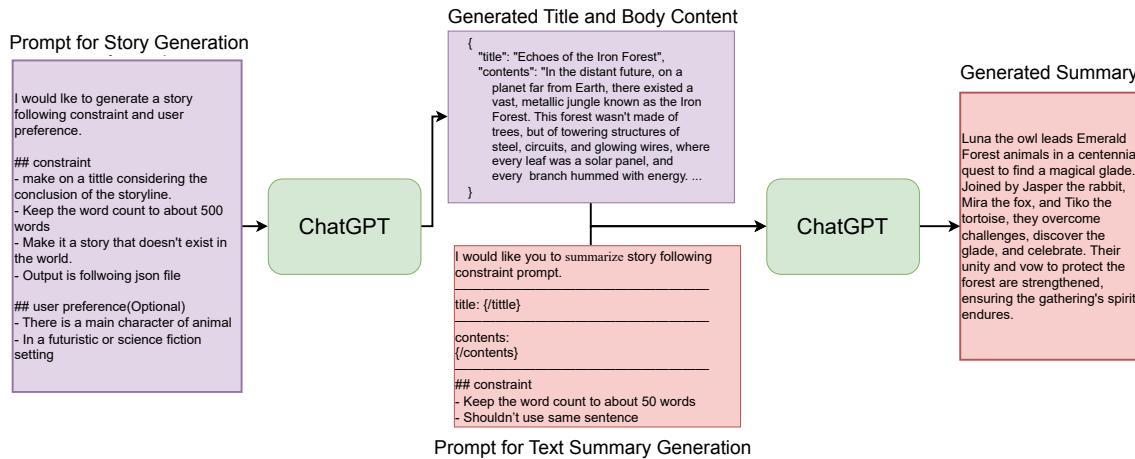


Figure 2: Architecture of the generation flow of the story text summary using LLMs (ChatGPT). The input consists of the generated story from the story generation phase and constraint prompts, which guide the summary generation. The constraint prompts control parameters such as word count, ensuring that the summary is concise and adheres to the specified length and content requirements for effective summarization.

enabled significant advancements in visual content creation, particularly in user interface design and interactive experiences [19, 39].

Parallel to LLMs, diffusion-based image generation models have emerged, introducing new possibilities for visual content creation in HCI [19, 39]. These models have facilitated advancements in user interface design and interactive experiences. However, their adoption in educational contexts remains limited, primarily due to the labor-intensive nature of prompt engineering [19]. Recent research has begun addressing these challenges by integrating text and image generative AI (T+IGenAI). Leveraging LLMs to generate prompts for image generation reduces user burden and enhances creative workflows [26]. Studies have explored using LLMs to assist in video editing [38] and facilitating collaborative creation between AI and humans [9]. Additionally, Generative AI has been employed to create fairy tales that combine textual and visual elements to evoke emotional responses [12].

Despite these advancements, there is a noticeable gap in research demonstrating the practical effectiveness of Generative AI in real-world educational applications. Our research aims to fill this gap by leveraging text and image generation AI to create visually enriched educational content. By generating content aligned with individual students' interests and preferences, we seek to increase engagement and motivation in the learning process.

2.2 Human Memory and Comprehension Augmentation

Understanding human memory and comprehension is crucial for enhancing educational outcomes. Research indicates that both textual and visual information play critical roles in memory retention, and combining these modalities can enhance learning.

Textual information involves complex cognitive processes. The presentation of information significantly influences reading comprehension [35] and memory retention [7]. For instance, coherent and well-structured text aids in better understanding and recall [17].

Elaborative encoding, which entails processing the meaning of information and linking it to existing knowledge, improves memory retention [6]. Visual information is processed differently and often has a substantial impact on memory and comprehension. According to Paivio's Dual Coding Theory, information encoded both verbally and non-verbally enhances comprehension [24]. This is supported by the picture superiority effect, where images are remembered better than words [22]. Shepard demonstrated that people could recall images accurately even after long delays, indicating robust visual memory [32]. Integrating text and visual information can significantly enhance memory retention. Mayer and Moreno found that combining words and pictures leads to better learning outcomes than using either alone [21]. This multimedia learning effect is particularly strong when visuals directly relate to the text, providing contextual support that aids understanding and retention. Carney and Levin showed that pictorial illustrations can improve memory for textual information by providing visual anchors that assist in retrieval [4]. Personal interest and the relevance of information also influence memory retention. Individuals are more likely to remember personally interesting or relevant information. Renninger and Hidi found that interest enhances cognitive processing, leading to better memory retention [29]. Similarly, Schiefele demonstrated that students interested in a topic recalled more information and had better comprehension than less interested peers [30].

Our research leverages these insights by integrating AI-generated supplemental content to enhance educational materials. By combining textual summaries and relevant images generated by AI, we aim to improve memory retention and comprehension, tailored to individual learner preferences.

3 METHODOLOGY

This study is structured around two key phases: the Text Generation Phase (TGP) and the Image Generation Phase (IGP). The

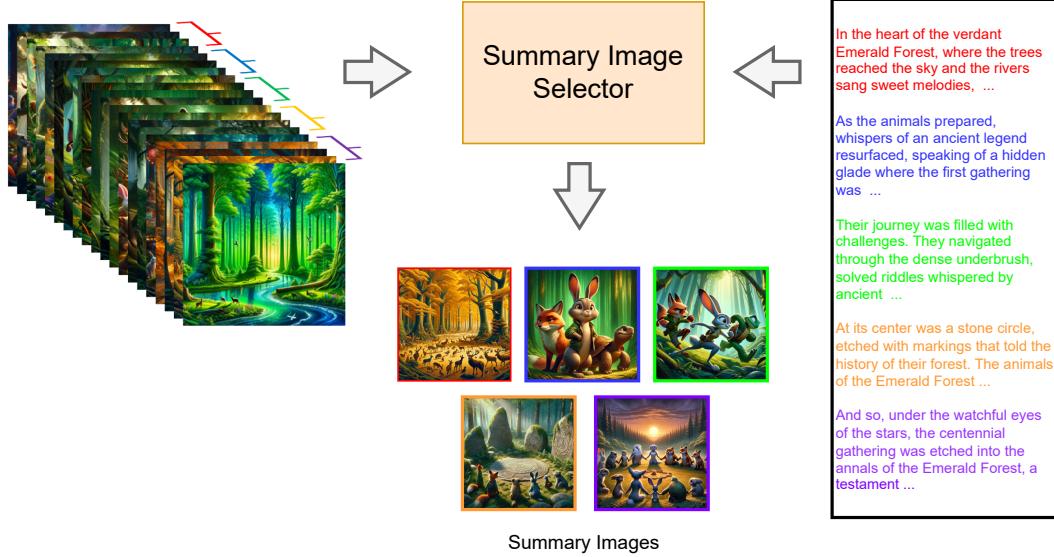


Figure 3: Architecture of the selection flow of the summary image selector. The input includes the story and the generated images, which are processed to select five key summary images. The text and images are segmented and fed into the Summary Image Selector to calculate the highest similarity score in each segment, which is chosen as the summary image.

TGP comprises two main components: story generation and text-summary generation. The IGP, on the other hand, includes the sentence-image generation and the summary-image selector. In the following sections, we explain the processes and methodologies involved in each phase.

3.1 Approach for the Text Generation (TGenAI)

Inspired by the StoryPrompt [8], we utilized the ChatGPT to generate story datasets. The ChatGPT is an advanced large language model based on GPT-3 [40], designed to generate coherent and contextually relevant text across a wide range of domains, using a combination of supervised learning and reinforcement learning from human feedback (RLHF) [23]. Firstly, we create the story within the Story Generation section. Once the story is generated, it is distilled into a concise summary in the Summary Generation section. This summary not only provides a quick reference for the story but also plays a crucial role in guiding the subsequent Image Generation phase, ensuring that the visual content accurately reflects the core elements of the story.

3.1.1 Story Generation. Figure 2 shows the input for this section consists of instruction, constraint, and preference prompts, and the output is a generated story that adheres to these parameters. The constraint prompt imposes specific limitations, such as word count restrictions and the requirement that the generated content must be completely original, ensuring no prior existence in any form. The preference prompt allows story customization based on user-defined elements, including preferred animals (e.g., rabbits, foxes) and story genres (e.g., adventure, science fiction). Users can also provide the title of a favorite story to guide the generation of a personalized story. Following the generation of the story, morphological and dependency parsing were conducted to analyze the

grammatical structure and relationships within the text. These analyses ensure that the visual elements generated in the subsequent phase accurately represent the story, maintaining consistency in character and thematic elements.

3.1.2 Summary Generation. The input for this section consists of the generated story from Section 3.1.1 and constraint prompt, and the output is a concise summary. The constraint prompts control of various aspects of the summary generation, including word count limitations, ensuring that the summary is concise and aligned with the specified length and content requirements.

3.2 Approach for the Image Generation (IGenAI)

Firstly, we create the story image from each sentence from the generated story. Then, the generated images input the Summary Image Selector to obtain the summary images.

3.2.1 Story Image Generation. In this section, we use the diffusion-based text-to-image model DALL-E [28] to generate highly detailed and contextually relevant images based on the given prompts. As shown in Figure 1, the inputs typically include the generated story and the preceding image to ensure stylistic continuity. In the initial image generation process, where no prior image exists, only the textual input is used as the basis for generating the first image in the sequence. The aim is to generate a series of images that match the sentences and maintain a consistent style throughout.

First, the generated story was segmented into individual sentences. Each sentence, the results of the dependency parsing, and the summary were used as input to generate corresponding images. This process involves sentence-to-image generation, where

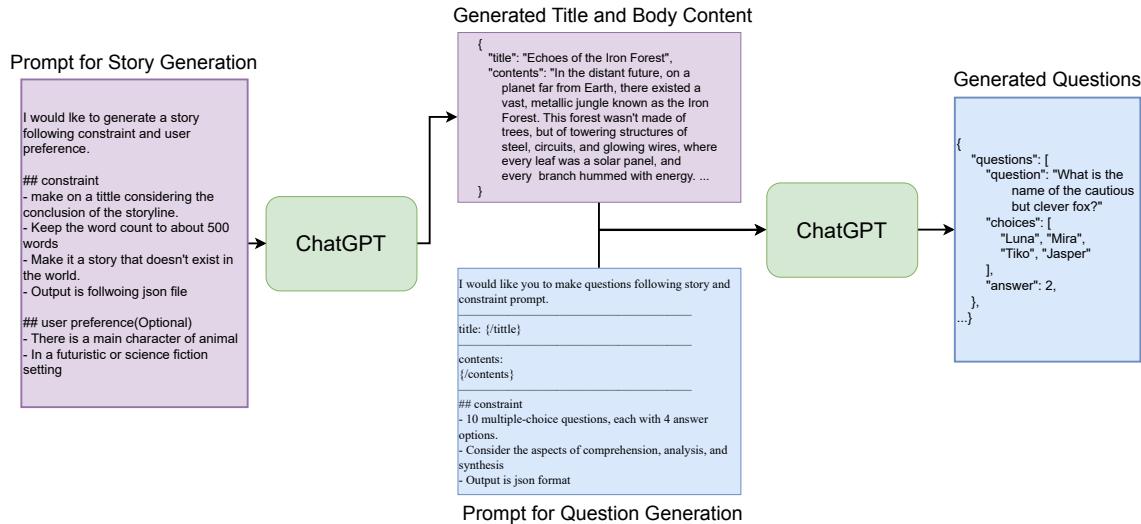


Figure 4: Architecture of the generation flow of the questions using LLMs (ChatGPT). The input is generated story from the story generation phase to tailor questions to align with the story content and constraint prompts. The prompts define question types, such as multiple-choice or open-ended, and determine the focus areas like numerical values or narrative comprehension, ensuring the output is formatted appropriately for further use.

the dependency parsing results - including character and style information - are continuously integrated into the prompts. This ensures that the generated images consistently reflect the narrative coherence and character continuity. Furthermore, a reference image from the previous output was incorporated as input to maintain visual consistency across the sequence of images. By including the story summary, the model could generate images that grasped the overall storyline and conclusion, enhancing the narrative's visual representation. We carry out this process regressively to obtain a series of images.

3.2.2 Summary Image Selection. In this section, we use the CLIP to select the summary images. The CLIP (Contrastive Language-Image Pre-training) [27] is a model that aligns images and text in a shared embedding space. By jointly training on a large dataset of images paired with textual descriptions, CLIP learns to associate textual and visual information effectively. It allows it to evaluate the similarity between a text and an image, making it particularly useful for tasks that require matching or retrieving images based on textual input.

As illustrated in Figure 3, this section requires the story and the generated story images as inputs to select five summary images as outputs. The story was divided into five segments based on the maximum token limit of the CLIP text encoder, ensuring that each segment is processed within the model's optimal capacity for accurate similarity calculation. The corresponding set of images generated in Section 3.2.1 was also segmented according to their respective sentences. Each segmented text and image was input into the CLIP encoder to obtain sentence and image vectors. The cosine similarity between these vector spaces was calculated, and the image with the highest similarity score for each segment was selected as the summary image for that part. The summary image

was determined as follows:

$$\text{CLIP-S}(c_i, v_j) = w \cdot \max(\cos(c_i, v_j), 0)$$

where $w = 2.5$, $\cos(c_i, v_j)$ denotes the cosine similarity between the textual embedding c_i and the visual embedding v_j . The weight $w = 2.5$ was chosen based on the findings of CLIP-Score [13], demonstrating that this value optimizes the balance between text-image alignment and visual representation. This weight emphasizes the highest similarity scores, ensuring that the most representative images are selected as summary images. To select the summary image for each segment i , we compute:

$$\text{summary_image}_i = \arg \max_j (\text{CLIP-S}(c_i, v_j)) \quad (i = 1, \dots, 5, j = 1, \dots, \frac{n}{5})$$

Based on the highest vector similarity, we select the most representative image as a summary image for each segment.

3.3 Preparation of Questions for Reading Comprehension Evaluation

Figure 4 shows the overall generating questions pipeline. We focus on generating questions that align with the story and constraints prompt, ensuring that the questions are tailored to the content and parameters of the generated story. The constraint prompts define the types of questions to be generated, including multiple-choice, open-ended, or fill-in-the-blank formats. Additionally, the prompts determine the specific focus of the questions, such as assessing numerical values, proper nouns, or broader narrative comprehension. Furthermore, the constraint prompts allow users to specify the desired output format and structure. This ensures the generated questions are organized in a suitable file type or format for subsequent use.

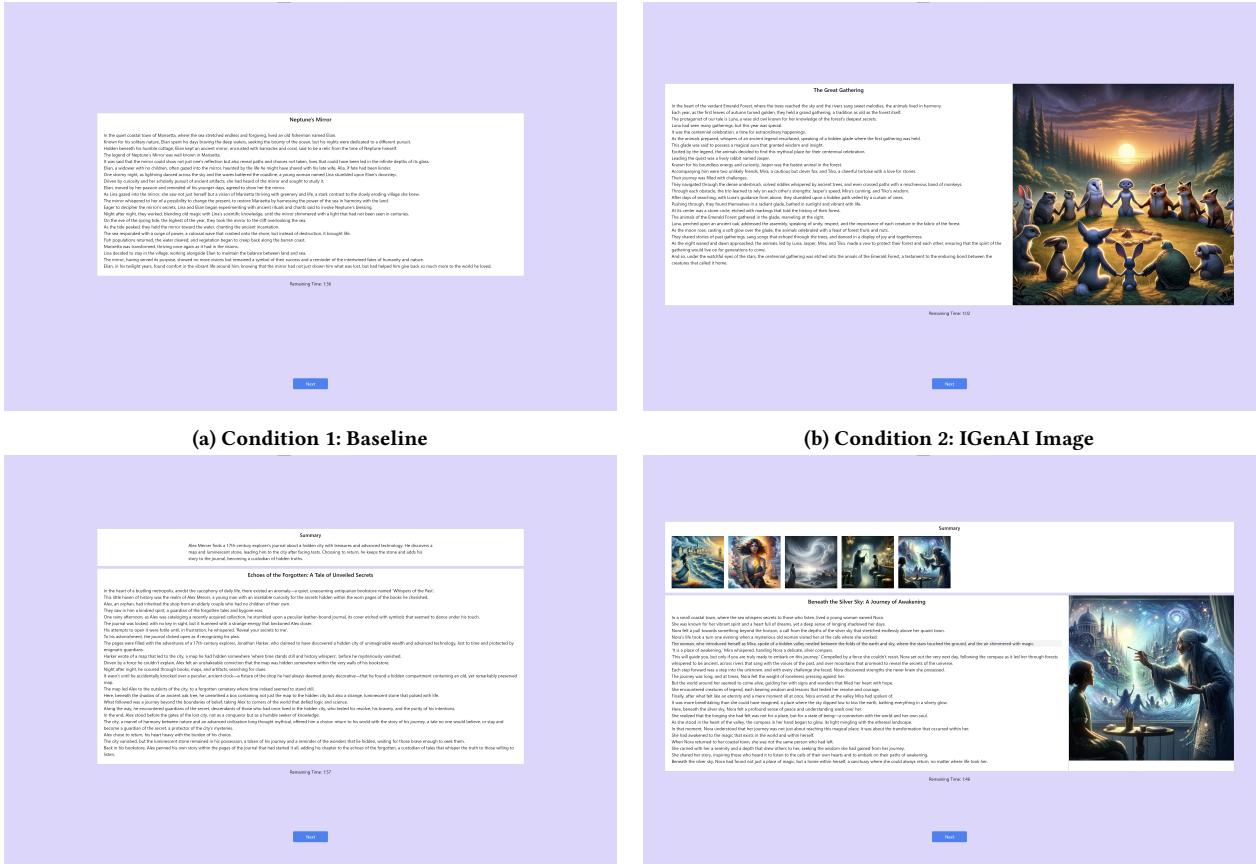


Figure 5: User interface of the web application showing the four reading conditions: (a) “Baseline”, (b) “IGenAI Image”, (c) “TGenAI summary”, and (d) “IGenAI Summary”.

ID	Condition	Detail
C1	Baseline	Participant read a document without any Generative AI augmentation.
C2	IGenAI Image	Participant read a document with Generative AI image augmentation.
C3	TGenAI Summary	Participant read a document with Generative AI text summary augmentation.
C4	IGenAI Summary	Participant read a document with Generative AI image summary augmentation.

Table 1: List of four reading conditions prepared in this study.

3.4 Web Application

In this section, we explain how our web application was implemented for the user study. First, we explain the system workflow of this application. Then, we explain the log metrics collected using the application.

3.4.1 Order of Reading Conditions. One story was randomly selected and designated as the fixed story to assess participants’ English language proficiency. The remaining three stories were divided into six distinct combinations ($3! = 6$), each assigned to one of the following conditions C2-C4 presented in Table 1. To present the content consistently and visually organized, an application was

developed using React and Node.js, as depicted in Figure 5. Upon initiating the task, the content for the base condition is centrally displayed on the screen. In the C2 condition, the text is presented on the left side of the screen, with the corresponding image displayed on the right. In the C3 condition, the time limit was set based on the length of the story, not the summary, to maintain a consistent information load. In the C4 condition, the image is positioned above the text.

3.4.2 User Interface of IGenAI. In the IGenAI page, before the reading task, participants were informed that hovering their mouse over a particular section of the text would trigger the generation of an image related to that specific content. As participants begin reading,

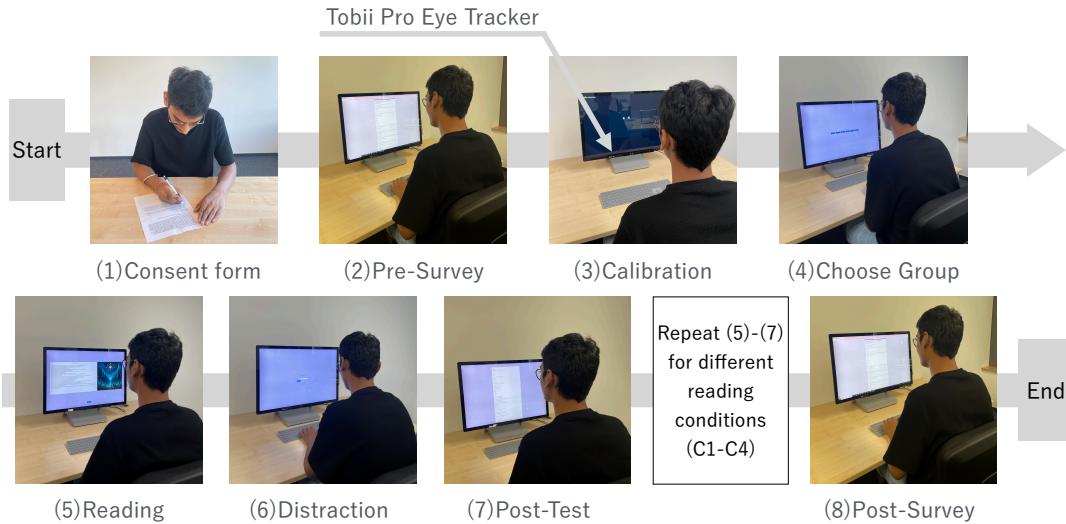


Figure 6: Experiment workflow. Calibration refers to an eye-tracker, the process of estimating the geometric characteristics of a subject’s eyes. The post-reading test provides ten questions for evaluating reading comprehension and memory retention of the provided reading conditions.

whenever they hover over a sentence or phrase, the application immediately generates and displays the corresponding image on the right side of the screen. If the participant moves the cursor away from the text, the last hovered image remains on the screen, ensuring continuous visual support throughout the reading process. This interactive feature allows participants to receive visual cues in real-time, enhancing their comprehension by providing immediate visual context for their reading text.

3.4.3 Reading Time Limit. The application is programmed to automatically transition to the next page once the allotted time limit is reached. Reading times were intentionally set to a faster rate of 250 words per minute [34], to challenge participants to process and retain information under a more demanding time constraint.

3.4.4 Log Metrics. In this study, we systematically collected log data and gaze behavior during the reading tasks.

- Duration: The time taken (in seconds) to read the text and complete the accompanying questions.
- Group Number: The number of groups the participants are.
- Story Index: The index of the story the participants read.
- Distraction Score: The total number of correct responses provided by participants in the distraction task.
- Number of Correct Answers: The total number of correct responses provided by participants.

3.5 Eye-Tracking Data Processing

3.5.1 Fixation. We extracted fixations by grouping nearby eye-tracking gaze points using the Identification by Dispersion-Threshold algorithm, as described by Buscher et al. [3]. We identify a new fixation when nine consecutive eye-tracking gaze points are detected within 50 pixels of each other. With our eye-tracker sampling at 90 Hz, the nine-point threshold sets the minimum fixation duration to 100 ms. The 50-pixel threshold ensures that the gaze points fall

within the same region of interest. To account for measurement noise and small eye movements such as microsaccades, the dispersion threshold is increased to 80 pixels for subsequent points. Thus, gaze points are added to the fixation one at a time as long as they fall within the area determined by the threshold. If a gaze point does not meet the requirement, the algorithm resets and begins counting a new set of nine gaze points to identify a new fixation.

3.5.2 Areas of Interest (AOI) and AOI Ratio. The calculated fixation duration and coordinates are used to extract the areas of interest (AOI) [16]. As shown in Figure 5, the reading content, generated image, and summary areas are held constant in our user interface. Our approach allows us to estimate whether the fixation is on the reading content, a generated image, or a summary. Using the coordinate of the fixation, we apply a threshold to separate the AOI in each content. Instead of using a fixation duration to evaluate the reading behavior, we use an AOI ratio. As explained in Section 3.4.3 the reading time duration differs between the document’s word count. Therefore, our experiment focuses on calculating the AOI ratio between the reading content area and the generated image or summary area.

4 DATA COLLECTION

In this section, we introduce details about the participants’ demographic information, experiment setup, and the experiment procedure.

4.1 Participants

This study recruited 24 participants from diverse national backgrounds, including Eastern Europe, South Asia, East Asia, the Middle East, South America, and North America. The participants, aged between 21 and 31 years with an average age of 27, were either

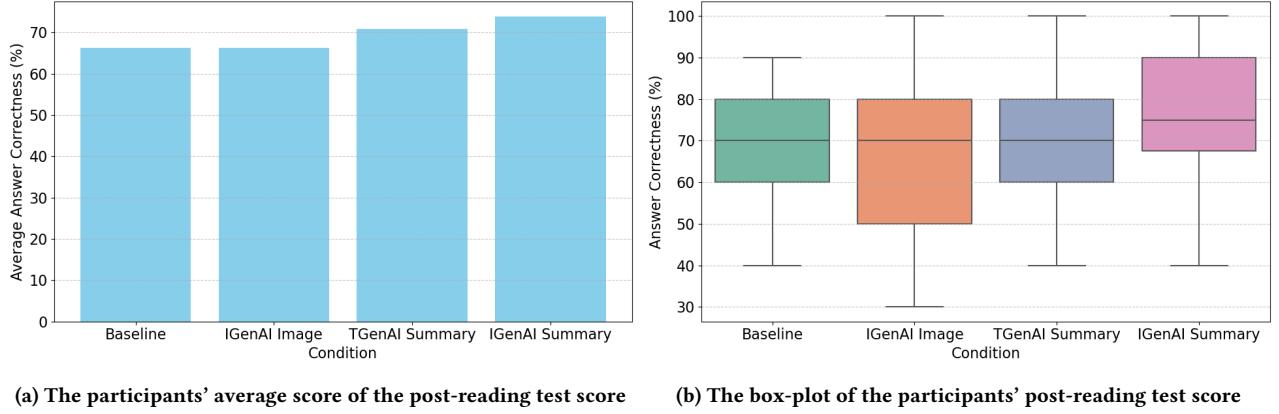


Figure 7: The comparison of the reading comprehension and memory retention using participants' post-reading test score results in all reading conditions (C1-C4).

Table 2: List of questions in the pre-survey.

Pre-Survey Questions	
Q1	Which group are you?
Q2	What is your name?
Q3	What is your age?
Q4	Where are you from? (Nationality)
Q5	What is your gender?
Q6	What is your occupation?
Q7	How long have you been using English?
Q8	What do you think about your English skills?
Q9	Are you familiar with LLMs?
Q10	Are you familiar with IGMs?

university students or professionals based in Germany. The General Data Protection Regulation (GDPR) required informed consent from all participants before the experiment. The participants were then assigned to six groups in a fair and unbiased manner, each tasked with solving problems under varying conditions.

4.2 Experiment Setup

During the story creation process, the following preference prompts were employed: by setting the configuration to nothing, animal, SF, and adventure, we ensured the diversity of the stories. Additionally, to avoid biases stemming from participants' prior knowledge, we generated entirely new stories that do not exist in the real world, thereby enabling fair comparisons. Each story was designed to be approximately 500 words in length. In the question creation phase, the preference prompts were set to generate multiple-choice questions. The questions assessed comprehension, memorization, and synthesis, comprehensively evaluating the participants' reading tasks. The output was formatted as a JSON file to facilitate data processing. Finally, for the summary creation phase, preference prompts generated summaries that were approximately 50 words in length.

Table 3: List of questions in the post-survey.

Post-Survey Questions	
Q1	To what extent are you familiar with Large Language Models (LLMs)?
Q2	To what extent are you familiar with Image Generation Models (IGMs)?
Q3	Which textbook (reading condition) did you find most interesting?
Q4	Which textbook was most helpful in aiding your memorization or in solving questions?
Q5	How did you perceive the allocated time for reading?

4.3 Experiment Procedure

Figure 6 shows our experiment's overall experiment setup and workflow. The Tobii Pro eye-tracker with a sampling rate of 90Hz, a remote device with an academic license that records eye movements, is mounted in the Microsoft Surface Studio 1. All participants used identical desktop computers equipped with Microsoft Surface Studio 1. Gen with 637.35×438.90 mm screen size, ensuring each story could be displayed on a single screen. The experiment was conducted in a quiet room. This controlled environment ensured all participants experienced the experiment under the same conditions. Participants first answered the consent form to confirm their participation in our data collection. The consent form includes informing the subjects about data protection and compliance with the GDPR.

Once the participants confirmed their participation in this experiment, they answered the pre-survey. Table 2 shows a list of questions asked in the pre-survey. As explained in Section 3.4.1, we divided the participants into six groups to avoid bias in performance due to the difficulty of each story in this experiment. Hence, we asked participants to indicate which group they would be involved in pre-survey Q1. The participant answered the remaining questions as appropriate. After the pre-survey, participants worked on Tobii Pro eye-tracker calibration. Participants look at the dots on the screen to estimate the geometric characteristics of a participant's

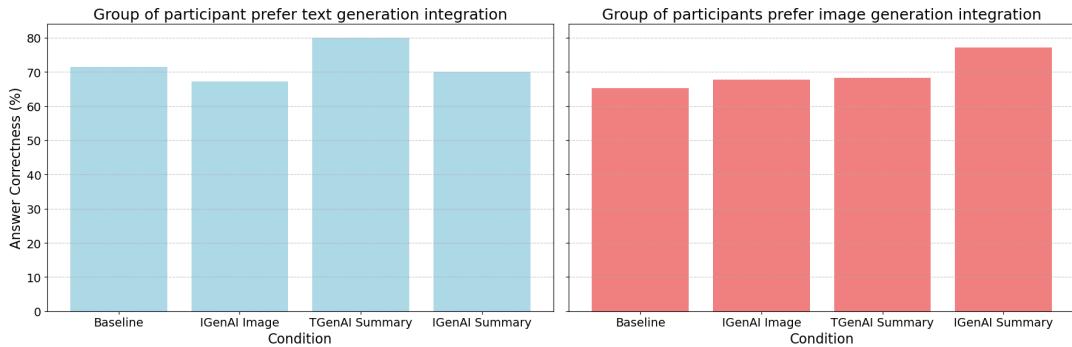


Figure 8: The comparison of post-reading test scores in solving questions between readers preferring text or image generation model.

eyes. Once calibration is done, we start eye-tracking data recordings. Participants then access the webpage¹, our experiment web application. The first page shows the selection of the group number, and the participants choose the one required by the experiment conductor.

Participants read four stories under different reading conditions (C1-C4). Each group read the stories in a specific sequence to control for order effects. The combination of story and reading conditions will vary from group to group. Participants were informed before the reading time limit to standardize the time spent on each story. The time limit for each story was adjusted according to the word count as explained in Section 3.4.3. Following the Atkinson, after reading each story, participants engaged in a distraction task [2] designed to clear their short-term memory and minimize any immediate recall effects. This task involved solving fundamental arithmetic problems for one minute. Once the distraction task is over, the participant answers ten post-reading tests related to the story the participant read. Participants then repeat the reading task with different stories and conditions (C2-C4). Once the participant completes all reading conditions (C1-C4), finish working with the web application.

Lastly, participants worked on the post-survey, whose questions are listed in Table 3. We ask questions that can qualify the subjective feedback of which textbook (reading condition) supports participants' reading comprehension.

5 RESULT AND DISCUSSION

In this section, we present the results of our study and discuss their implications. We begin by examining overall reading comprehension across different conditions, followed by an analysis of the relationship between participants' preferences and their reading performance. We then explore eye movement patterns to understand how participants interacted with the GenAI materials.

5.1 Overall Reading Comprehension Across Conditions

Figure 7 shows the average post-reading test scores under different reading conditions. The *IGenAI Image* condition (text with

AI-generated images) yielded an average score 1.25% higher than the *Baseline* (text-only), suggesting a slight improvement in comprehension when images are included. However, the *IGenAI Image* condition exhibited greater variability in scores compared to the consistently high performance in the *Baseline*, indicating that while some participants benefited from the images, others did not experience the same improvement.

Both GenAI summary conditions (*TGenAI Summary* and *IGenAI Summary*) significantly outperformed the *Baseline*, highlighting that providing summaries—whether text-based or image-based—enhances reading comprehension. Notably, the *IGenAI Summary* condition showed more variability in scores than the *TGenAI Summary*, suggesting that image summaries are highly effective for some readers but less helpful for others.

These findings indicate that adding images as supplementary material introduces more variability in performance compared to adding text summaries. While images can enhance comprehension for some participants, they may also lead to inconsistent outcomes due to individual differences in processing visual information.

5.2 Relationship Between Participants' Preferences and Reading Comprehension

Figure 8 illustrates the relationship between participants' preferences for learning material formats and their corresponding post-reading test scores. Participants were grouped based on their preferred medium: text generation (*TGenAI*) and image generation (*IGenAI*). This division allowed for a detailed analysis of how different types of supplementary materials impacted their comprehension and retention.

In the text-preference group (*TGenAI*), the highest average test scores were achieved under the *TGenAI Summary* condition, demonstrating the benefit of providing a text-based overview. The *Baseline* condition, which consisted of text-only content, yielded the second-highest scores. In contrast, participants in this group performed noticeably worse under the *IGenAI Image* condition, and their lowest scores were recorded under the *IGenAI Summary* condition. This suggests that visual summaries and images were less effective for participants favoring text-based materials.

Conversely, in the image-preference group (*IGenAI*), the *IGenAI Summary* condition produced the highest test scores. Participants

¹<https://generative-ai-textbooks.netlify.app/>

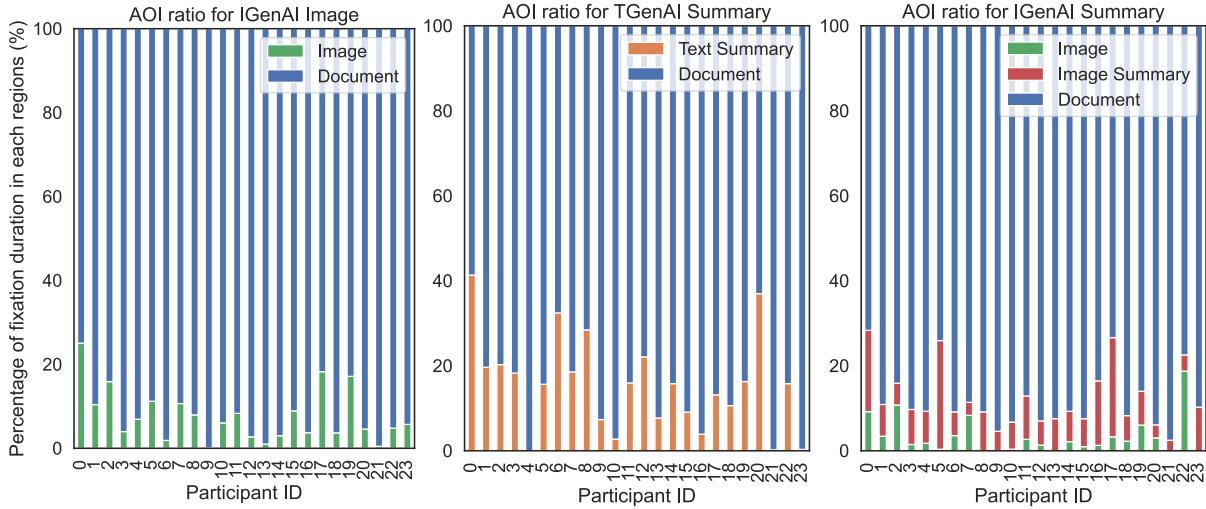


Figure 9: Comparison of the AOI ratio calculated by fixation duration for each participant in three different reading conditions (C2-C4). The “Document” represents the main reading text content. “Image”, “Text Summary”, and “Image Summary” represent the AOI ratio while looking at each area in the user interface prepared by IGenAI Image, TGenAI Summary, and IGenAI Summary.

in this group also performed relatively well under the *IGenAI Image* and *TGenAI Summary* conditions. However, their lowest scores were observed in the *Baseline* condition, indicating that the absence of visual aids hindered their comprehension and learning outcomes.

These findings highlight the importance of aligning supplementary educational materials with learners’ preferences. Participants favoring text generation achieved higher scores with TGenAI materials, particularly when supplemented with text summaries. Similarly, those inclined toward image generation performed best with IGenAI materials, benefiting from both image summaries and standalone images. These results suggest that tailoring educational content to individual preferences—textual or visual—not only improves engagement but also enhances comprehension, retention, and overall learning outcomes.

By leveraging personalized supplementary materials, this study underscores the potential of cognitive augmentation in education. Customizing content delivery based on learners’ preferences can foster a more effective and enjoyable learning experience, paving the way for improved academic performance and deeper intellectual engagement.

5.3 Eye Movement Patterns and Reading Behavior

We analyzed eye-tracking data to gain deeper insights into how participants engaged with the GenAI materials.

5.3.1 Evaluation of AOI Ratio on Different Reading Conditions. Figure 9 presents the AOI ratios, calculated by fixation duration, for each participant across three reading conditions (*IGenAI Image*, *TGenAI Summary*, and *IGenAI Summary*). The results reveal distinct reading behaviors among participants. For example, *Participant 17* demonstrated a strong preference for visual information, spending significant time viewing images in both IGenAI conditions.

In contrast, *Participant 20* focused predominantly on the *TGenAI Summary*, indicating a preference for textual material. *Participant 0* engaged extensively with all forms of supplementary information, while *Participants 4* and *23* showed disinterest in the *Text Summary*. *Participant 9* largely ignored visual information, focusing on text, and *Participant 21* exhibited minimal engagement with any GenAI material. These observations illustrate that individuals process information differently, reflecting diverse cognitive strategies. GenAI summaries were generally more engaged than single images, possibly due to their comprehensive nature. This underscores the importance of acknowledging individual differences in educational design and the potential benefits of personalized learning materials.

5.3.2 Impact of Summaries on Reading Engagement. Figure 10 compares the gaze scan paths of a participant under different reading conditions. When supported by the *TGenAI summary* (Figures 10a and 10b), the participant engaged with a wider range of document regions than in the baseline condition, suggesting that the textual summary facilitated more efficient navigation.

Including summaries significantly improved performance, as illustrated in Figure 7. The presence of a summary allowed readers to grasp the story’s overall structure early on, enabling quicker comprehension of subsequent content. Further analysis shows that both text and image summaries enhanced performance, though their impact varied based on reading speed. Text summaries required additional reading time, while image summaries provided an almost instant understanding of the story’s flow, freeing up time for the main text and enhancing comprehension.

These results suggest that summaries function as an effective pre-reading tool for cognitive augmentation. In time-sensitive scenarios like this study, visual summaries outperformed text summaries due to their rapid processing advantage. The *Summary Image Selector* developed in this research holds promise for improving AI-generated



Figure 10: Comparison of the gaze scan path for different reading conditions for “Baseline”, “TGenAI Summary”, “IGenAI Image” and “IGenAI Summary”.

textbooks, particularly for speed-reading tasks, thereby fostering more efficient learning experiences.

5.3.3 Verbal vs. Visual Preference Learners. Based on participants’ post-reading test scores and self-reported preferences, we observed a tendency suggesting that individual learning preferences might influence reading comprehension. Some participants, termed *Verbal Preference Learners*, appeared to process information more effectively through language and text, while others, *Visual Preference Learners*, excelled when information was presented visually.

To explore the implications of these learning styles, we conducted further analysis. While we cannot definitively claim a statistical correlation, the observed patterns indicate that personalization could enhance learning outcomes. We analyzed gaze heatmaps to illustrate these tendencies. As shown in Figure 11, *Participant 19*, who preferred text generation, progressed through the text with minimal engagement with images, even when visual elements were present. In contrast, participants who preferred image generation frequently alternated their gaze between text and images, indicating active integration of visual information.

These findings highlight the importance of incorporating both verbal and visual elements into educational materials to accommodate diverse learner preferences. Our model demonstrates that personalization is possible by providing supplementary materials that align with individual cognitive styles. By tailoring educational content to match learners’ preferences, we can potentially enhance comprehension and retention across varied student populations.

6 LIMITATIONS AND FUTURE DIRECTIONS

Our study indicates that AI-generated supplementary materials can significantly enhance post-reading test scores. However, several limitations remain, offering multiple avenues for further investigation.

Computational Bottleneck in Image Generation. The computational intensity of current diffusion-based models (e.g., DALL-E) poses scalability challenges, especially for real-time or large-scale educational contexts. Even smaller class settings may struggle with on-demand generation when bandwidth or hardware resources are limited. Future work might explore more efficient pipelines, hardware optimizations (e.g., GPU clusters), or specialized smaller models that maintain quality while reducing latency.



(a) Gaze fixation heatmap of the verbal preference learner (PID=9)

(b) Gaze fixation heatmap of the visual preference learner (PID=0)

Figure 11: The comparison with the reading behavior between verbal and visual preference learner. In the post-survey (Q3 and Q4), verbal preference learner is the sample of participants who chose “text-generation” as a preference for the reading condition, and visual preference learners chose “image-generation” as a preference for the reading condition.

Domain-Specific Complexity. We focused on relatively simple narratives, for which the visual representations are straightforward. More specialized fields—such as chemistry, physics, or mathematics—require domain-specific visuals (e.g., chemical structures, equations) that exceed the capabilities of general-purpose generative models. Techniques like fine-tuning, advanced prompting, or hybrid approaches that combine symbolic knowledge with generative AI could improve the precision and realism needed in these complex subjects.

Cognitive Load and User Interface. Although supplementary images and summaries aided comprehension, additional on-screen elements led to more frequent gaze shifts (Figures 10c and 10d), which some participants described as distracting. Future research might evaluate adaptive UI layouts—such as overlays or collapsible panels—that appear only when needed. Such adaptive interfaces could balance the benefits of multimodal information against the risk of overloading the learner’s visual attention.

Potential Bias in Participant Demographics and AI Familiarity. Our participants, largely young adults with prior experience using Large Language Models (LLMs) and Image Generation Models (IGMs), may not represent broader or less tech-savvy populations (see Figure ?? in the supplementary material). Individuals with different cultural backgrounds, age ranges, or language proficiencies may respond differently to AI-augmented materials. Recruiting more diverse user groups—including older adults, K-12 students, or non-native English speakers with minimal AI exposure—can deepen understanding of how demographic factors modulate the effectiveness of generative textbooks.

Reliance on AI-Generated Stories and Questions. We employed LLM-generated narratives and comprehension questions to avoid relying on participants’ prior domain knowledge. However, these generated materials do not necessarily reflect the rigor or pedagogical structure of professionally authored textbooks or standardized tests. AI-generated questions may overemphasize factual recall at the expense of deeper inference. Future research can integrate established texts and validated

question banks to better assess how generative AI translates to real-world educational settings.

Shallow Eye-Tracking Measures. Our study mainly considered fixation durations and areas of interest (AOIs), which capture broad viewing patterns. Additional psychophysiological or behavioral data, such as pupil dilation (an indicator of cognitive load), EEG signals (for workload), or changes in reading speed over time, might offer more nuanced insights into how AI-generated content shapes comprehension and engagement. Further work combining these measures with eye-tracking could refine our understanding of the mechanisms behind AI-enhanced reading.

7 CONCLUSION

This study demonstrates that incorporating AI-generated supplementary materials—such as images, text summaries, and image summaries—significantly enhances cognitive augmentation. Our findings indicate that learners’ preferences affect reading behavior and post-reading scores, challenging the traditional one-size-fits-all approach to educational materials. Enhancing traditional text-based materials with AI-generated content can improve memory retention and comprehension. The study recruited 24 participants, and verified that integrating AI-generated supplementary materials significantly improved learning outcomes, increasing post-reading test scores by 7.50%. As AI technology continues to evolve, integrating it into educational tools represents a powerful strategy for fostering learner engagement and understanding. This research highlights the potential of generative AI to create personalized educational experiences that enhance human intellect and cognitive augmentation. Our study bridges gaps in traditional education and encourages the broader adoption of AI in personalized learning environments.

ACKNOWLEDGMENTS

This work is partially supported by JSPS KAKENHI Grant Number 24K02962.

REFERENCES

- [1] Farhan Ali, Doris Choy, Shanti Divaharan, Hui Yong Tay, and Wenli Chen. 2023. Supporting self-directed learning and self-assessment using TeacherGAIA, a generative AI chatbot application: Learning approaches and prompt engineering. *Learning: Research and Practice* 9, 2 (2023), 135–147.
- [2] Richard C Atkinson. 1968. Human memory: A proposed system and its control processes. *The Psychology of Learning and Motivation* 2 (1968).
- [3] Georg Buscher, Andreas Dengel, and Ludger van Elst. 2008. Eye movements as implicit relevance feedback. In *CHI '08 Extended Abstracts on Human Factors in Computing Systems* (Florence, Italy) (*CHI EA '08*). Association for Computing Machinery, New York, NY, USA, 2991–2996. <https://doi.org/10.1145/1358628.1358796>
- [4] Russell N. Carney and Joel R. Levin. 2002. Pictorial illustrations still improve students' learning from text. *Educational Psychology Review* 14, 1 (2002), 5–26.
- [5] Sarah Clinch and Jamie A Ward. 2023. Augmented Cognition. *IEEE Pervasive Computing* 22, 3 (2023), 6–7.
- [6] Fergus I. M. Craik and Robert S. Lockhart. 1972. Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior* 11, 6 (1972), 671–684.
- [7] Passant Elagroudy, Sebastian Feger, and Albrecht Schmidt. 2022. A Model for Selecting Media Type of Memory Cues in Ubiquitous Prostheses. In *Proceedings of the Augmented Humans International Conference 2022* (Kashiwa, Chiba, Japan) (*AHs '22*). Association for Computing Machinery, New York, NY, USA, 313–316. <https://doi.org/10.1145/3519391.3524169>
- [8] Min Fan, Xinyue Cui, Jing Hao, Renxuan Ye, Wanqing Ma, Xin Tong, and Meng Li. 2024. StoryPrompt: Exploring the Design Space of an AI-Empowered Creative Storytelling System for Elementary Children. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–8.
- [9] Xianzhe Fan, Zihan Wu, Chun Yu, Fenggui Rao, Weinan Shi, and Teng Tu. 2024. ContextCam: Bridging Context Awareness with Creative Human-AI Image Co-Creation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–17.
- [10] Soheil Feizi, MohammadTaghi Hajiraghayi, Keivan Rezaei, and Suho Shin. 2023. Online Advertisements with LLMs: Opportunities and Challenges. *arXiv preprint arXiv:2311.07601* (2023).
- [11] André Pimenta Freire, Paula Christina Figueira Cardoso, and André de Lima Salgado. 2023. May We Consult ChatGPT in Our Human-Computer Interaction Written Exam? An Experience Report After a Professor Answered Yes. In *Proceedings of the XXII Brazilian Symposium on Human Factors in Computing Systems*. 1–11.
- [12] Lasse Harde, Lasse Jensen, Johan Krogh, Adrian Plesner, Oliver Sørensen, and Henning Pohl. 2024. The Generative Fairy Tale of Scary Little Red Riding Hood. In *Proceedings of the 2024 ACM International Conference on Interactive Media Experiences*. 129–144.
- [13] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. ClipScore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718* (2021).
- [14] Yihan Hou, Manling Yang, Hao Cui, Lei Wang, Jie Xu, and Wei Zeng. 2024. C2Ideas: Supporting Creative Interior Color Design Ideation with a Large Language Model. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–18.
- [15] Shoya Ishimaru, Nicolas Großmann, Andreas Dengel, Ko Watanabe, Yutaka Arakawa, Carina Heisel, Pascal Klein, and Jochen Kuhn. 2018. Hypermind builder: Pervasive user interface to create intelligent interactive documents. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*. 357–360.
- [16] Andy J King, Nadine Bol, R Glenn Cummins, and Kevin K John. 2019. Improving visual behavior research in communication science: An overview, review, and reporting recommendations for using eye-tracking methods. *Communication Methods and Measures* 13, 3 (2019), 149–177.
- [17] Walter Kintsch and Katherine A. Rawson. 2005. *Comprehension*. Blackwell Publishing.
- [18] Kai Kunze, Pattie Maes, Florian Floyd' Mueller, and Katrin Wolf. 2023. Cognitive Augmentation (Dagstuhl Seminar 22491). (2023).
- [19] Vivian Liu and Lydia B Chilton. 2022. Design guidelines for prompt engineering text-to-image generative models. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–23.
- [20] Julia M Markel, Steven G Opferman, James A Landay, and Chris Piech. 2023. Gpteach: Interactive ta training with gpt-based students. In *Proceedings of the tenth acm conference on learning@ scale*. 226–236.
- [21] Richard E Mayer and Roxana Moreno. 2003. Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist* 38, 1 (2003), 43–52.
- [22] Douglas L. Nelson, Valorie S. Reed, and John R. Walling. 1976. Picture superiority effect. *Journal of Experimental Psychology: Human Learning and Memory* 2, 5 (1976), 523–528.
- [23] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems* 35 (2022), 27730–27744.
- [24] Allan Paivio. 1991. Dual coding theory: Retrospect and current status. *Canadian Journal of Psychology* 45, 3 (1991), 255–287.
- [25] Hyanghee Park and Daehwan Ahn. 2024. The Promise and Peril of ChatGPT in Higher Education: Opportunities, Challenges, and Design Implications. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–21.
- [26] Savvas Petridis, Michael Terry, and Carrie J Cai. 2024. PromptInfuser: How Tightly Coupling AI and UI Design Impacts Designers' Workflows. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference*. 743–756.
- [27] Alex Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.
- [28] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International conference on machine learning*. PMLR, 8821–8831.
- [29] K. Ann Renninger and Suzanne Hidi. 2015. *The Power of Interest for Motivation and Engagement*. Routledge.
- [30] Ulrich Schiefele. 1991. Interest, learning, and motivation. *Educational Psychologist* 26, 3–4 (1991), 299–323.
- [31] Albrecht Schmidt. 2017. Augmenting Human Intellect and Amplifying Perception and Cognition. *IEEE Pervasive Computing* 16, 1 (2017), 6–10. <https://doi.org/10.1109/MPRV.2017.8>
- [32] Roger N. Shepard. 1967. Recognition memory for words, sentences, and pictures. *Journal of Verbal Learning and Verbal Behavior* 6, 1 (1967), 156–163.
- [33] Haotian Sun, Yuchen Zhuang, Lingkai Kong, Bo Dai, and Chao Zhang. 2024. Adapanner: Adaptive planning from feedback with language models. *Advances in Neural Information Processing Systems* 36 (2024).
- [34] Susanne Trauzettel-Klosinski, Klaus Dietz, IREST Study Group, et al. 2012. Standardized assessment of reading performance: The new international reading speed texts IREST. *Investigative ophthalmology & visual science* 53, 9 (2012), 5452–5461.
- [35] Pramod Vadiraja, Andreas Dengel, and Shoya Ishimaru. 2021. Text Summary Augmentation for Intelligent Reading Assistant. In *Proceedings of the Augmented Humans International Conference 2021* (Rovaniemi, Finland) (*AHs '21*). Association for Computing Machinery, New York, NY, USA, 319–321. <https://doi.org/10.1145/3458709.3459002>
- [36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
- [37] Craig Vear, Adrian Hazzard, Solomiya Moroz, and Johanna Benerradi. 2024. Jess+: AI and robotics with inclusive music-making. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–17.
- [38] Bryan Wang, Yuliang Li, Zhaoyang Lv, Haijun Xia, Yan Xu, and Raj Sodhi. 2024. LAVE: LLM-Powered Agent Assistance and Language Augmentation for Video Editing. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*. 699–714.
- [39] Zhipie Wang, Yuheng Huang, Da Song, Lei Ma, and Tianyi Zhang. 2024. PromptCharm: Text-to-Image Generation through Multi-modal Prompting and Refinement. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–21.
- [40] Genta Indra Winata, Andrea Madotto, Zhaojiang Lin, Rosanne Liu, Jason Yosinski, and Pascala Fung. 2021. Language models are few-shot multilingual learners. *arXiv preprint arXiv:2109.07684* (2021).
- [41] Kanta Yamaoka, Ko Watanabe, Koichi Kise, Andreas Dengel, and Shoya Ishimaru. 2022. Experience is the best teacher: Personalized vocabulary building within the context of Instagram posts and sentences from GPT-3. In *Adjunct Proceedings of the 2022 ACM International Joint Conference on Pervasive and Ubiquitous Computing and the 2022 ACM International Symposium on Wearable Computers*. 313–316.
- [42] Qi Yang, Sergey Nikolenko, Marlo Ongpin, Ilia Gossoudarev, Yu-Yi Chu-Farseeva, and Aleksandr Farseev. 2024. SOMONITOR: Explainable Marketing Data Processing and Analysis with Large Language Models. *arXiv preprint arXiv:2407.13117* (2024).
- [43] Ruibin Yuan, Hanfeng Lin, Yi Wang, Zeyue Tian, Shangda Wu, Tianhao Shen, Ge Zhang, Yuhang Wu, Cong Liu, Ziya Zhou, et al. 2024. Chatmusician: Understanding and generating music intrinsically with llm. *arXiv preprint arXiv:2402.16153* (2024).

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009