IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 6 December 2024, accepted 20 January 2025, date of publication 23 January 2025, date of current version 31 January 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3533076

RESEARCH ARTICLE

Img2Vocab: Explore Words Tied to Your Life With LLMs and Social Media Images

KANTA YAMAOKA [1], KO WATANABE [2,3], KOICHI KISE [4], (Member, IEEE), ANDREAS DENGEL [2,3], AND SHOYA ISHIMARU [4], (Member, IEEE)

[1] College of Engineering, Osaka Prefecture University, Osaka 599-8531, Japan
[2] Department of Computer Science, RPTU Kaiserslautern-Landau, 67663 Kaiserslautern, Germany
[3] Smart Data and Knowledge Services Department, German Research Center for Artificial Intelligence (DFKI GmbH), 67663 Kaiserslautern, Germany
[4] Graduate School of Informatics, Osaka Metropolitan University, Osaka 599-8531, Japan

Corresponding author: Kanta Yamaoka (kantayamaoka@gmail.com)

This work was supported in part by the Deutsche Forschungsgemeinschaft (DFG) International Call on Artificial Intelligence Learning Cyclotron, in part by the Japan Society for the Promotion of Science (JSPS) Fund for the Promotion of Joint International Research (Grant Number: 23KK0188), and in part by the Japan Student Services Organization (JASSO) Student Exchange Support Program. The work of Kanta Yamaoka was supported by German Research Center for Artificial Intelligence (DFKI GmbH) During the Project.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by DFKI and OMU under LeCycl project.

ABSTRACT Psychological studies highlight the importance of combining new knowledge with one's prior experience. Hence personalization for a learner plays a key role for vocabulary acquisition. However, this faces two challenges: probing a learner's experiences in their lives and crafting tailored material for every different individual. With the prevalence of visual social media, such as Instagram, people share their photos from favorite moments, providing rich contexts, and emerging generative AI would create learning material in an effortless fashion. We prototyped an online vocabulary exploration system, which displays a learner's selected photos from their Instagram along with a generated sentence using image recognition and a language model, GPT-3. The system lets a learner find new words that are strongly tied to their daily life with the approximated context. We carried out our within-subject design evaluation of the system with 23 participants with three conditions: contexts grounded with learner's Instagram photos, contexts grounded from general images, and text-only modality. From learners' recall task accuracy, we found that having a context grounded with a learner's social image allowed them to find difficult words to quickly learn than having context generated by someone's image, or text only modality—although this finding is statistically insignificant. The Zipf frequency comparisons revealed that generally having image-based context allowed learners to extract difficult vocabulary than having text-only context. We also discuss quantitative and qualitative results regarding participants' acceptance of the personalization system using their personal photos from social media. Generally, they reported positive impressions for our system such as high engagement. While our system prioritizes user privacy with opt-in data control and secure design, we explore additional ethical considerations. This paves the way for a future where personalized language learning, grounded in real-world experiences and generative AI, benefits learners.

INDEX TERMS Context-aware language learning, HCI, large language models, generative AI.

I. INTRODUCTION

Personalization of learning plays a vital role in effective language learning, particularly when acquiring vocabulary for daily conversation. First, applying new knowledge to learn with prior experience, stored as episodic memory,

The associate editor coordinating the review of this manuscript and approving it for publication was Thomas Canhao Xu.

is effective because it entails a more profound encoding process, according to psychology research [1], [2]. Second, if a word is closely tied to a learner's daily life, the person is likely to use or recall the word repeatedly, which might turn the word from a passive vocabulary to an active vocabulary [3]. An active vocabulary related to their lives helps a learner in terms of conversation in English in their daily lives. Third, learning words tied to their personal lives is enjoyable because the learners can relate to their interests or hobbies, keeping them engaged [4], [5] with their long-term commitment. Although personalizing vocabulary learning based on a learner's experience is desirable, this brings about two challenges: (i) sensing a learner's experience and (ii) creating material for each person.

First, sensing a learner's experience is challenging. This is especially the case in conventional language education in a classroom. In the past, teachers had few clues about such personal backgrounds other than asking questions in person, such as "What are your hobbies?" or "What did you do last weekend?." However, this approach is not feasible considering the effort or cost it requires for learners and teachers. Today, social media are widespread among the young generation [6], and people have started posting pictures to describe their experiences, thoughts, or emotions. These posts represent their experience. They take pictures, especially when they are impressed, touched, or excited about something in their personal lives. Hence, such photos can be a new data point for assessing a learner's experience in daily life. As a concrete example, Instagram[1] is promising for this direction. Hu et al. [7] found that the pictures users post on Instagram were classified into the following eight categories: "self-portraits, friends, activities, captioned photos (pictures with embedded text), food, gadgets, fashion, and pets." This variety in users' photos indicates that Instagram can be a good resource to capture a person's experience.

Second, creating material for each person is painstaking work for human teachers. Suppose we can gain sufficient information about the learners; personalizing based on it is another difficult task for language learning. However, recent advancements in large language models (LLMs) can potentially mitigate this issue. One of the impressive examples of LLMs is GPT-3, a text-to-text model announced by OpenAI in 2020 learned with 175-billion parameters [8] and publicly available as API since the end of 2021.[2] Newly emerging technology provides the capability of creating versatile text from instruction by text. Although LLMs research is ongoing among many researchers and still under improvement, integrating such LLMs into a language learning system to provide personalized material seems a promising direction to mitigate effort for learning personalization.

To mitigate the obstacles to sensing a learner's experience and providing personalized material for each language learner, we devised a system to combine social media pictures and LLMs. In our work, we set the primary research hypothesis "Displaying a picture taken by a learner and a related sentence helps the learner explore unknown words in a target language." Concretely, we aim to allow learners to find new words to learn in a context tied to their life experiences. Discovering new words by itself is insufficient. Vocabulary acquisition should focus on learning difficult words while they are tied to their lives. There are two aspects to consider for a *difficult word*: (a) A word is difficult to guess the meaning of by context such as visual information. (b) A word is difficult to learn after immediate exposure, requiring learning efforts. We can paraphrase the idea behind them: a word easily guessable by the context or easily learnable is less important. These two types of words can be a bottleneck to understanding or describing the things happening in life; therefore, they must be discovered.

With the research hypothesis in mind, we developed and evaluated our first-of-its-kind system displaying learners' images imported from a learner's Instagram and generated sentences provided by image recognition and GPT-3. This way, learners can actively find new vocabulary that approximates their lives and interests. Our experimental results from a large-scale study with 23 participants indicate the system's opportunities for prioritizing words to learn that are: (a) difficult to guess their meanings without context and (b) difficult to recall their meanings after an immediate distraction task. Our system provides stretched goals by word lists approximating a person's life.

Our work contributes to research communities of HCI and EdTech realm with the following aspects:

KC1 We prototyped a new vocabulary exploration system and paradigm, where its context is generated from a learner's real-world experience using social media photos and a text-generation model, mitigating traditional difficulties of personalized vocabulary learning as to probing a learner's life and content-creation efforts.

KC2 Our quantitative and qualitative study with 23 participants highlighted the system's benefits to find difficult words yet also tied to a learner's daily experience, measured by the extracted word frequency comparisons and accuracy comparisons from post test.

KC3 Our user studies reveal the learners' positive impressions, such as indicated by user engagement, toward a system grounding learning material to their social media while also leading to privacy implications. We discussed further implications, aligned with existing privacy framework, to suggest future improvement in technical and ethical aspects for better technology acceptance of our system and similar learning personalization systems we anticipate to emerge.

II. RELATED WORK

This section first reviews the literature about personal experience and learning, which is the basis of our proposed methods. Second, we review context-aware language systems

[1] https://www.instagram.com/

[2] https://openai.com/blog/api-no-waitlist/

that intend to personalize learning and illustrate their room for improvement. Third, we review emerging technology, the large language models (LLMs), and their current usage in education.

A. LEARNING AND PERSONAL EXPERIENCE

In this section, we review existing studies on the relationship between learning and personal experience. They are the fundamental for our idea to personalize learning based on a learner's Instagram posts.

In our daily life, we are exposed to many stimulus by visual, audio and other sensory stimulus with different modalities, which shapes our experiences. The realms of psychology and the cognitive neuroscience have clarified that such experiences were encoded into our memory [9], and new incoming information can be consolidated on top of such existing knowledge. [1], [10], [11] First, from theoretical perspective, Craik et al. [9] framed memory as depths, or levels, of perceptual processing in various modalities, audio, visual and olfactory information as well as verbal information, and such processes involve assocations with existing knowledge. Second, as an example from experimental perspectives, Bransford et al. [10] illustrated the relationship between a prior-knowledge and recall by experiments, where participants who were given cues, a prior knowledge, before they listen to passages showcased better recall and comprehension compared to other participants who were not exposed to such interventions. As another example, Rogers et al. have empirically shown "self-reference effect," which is about associating new content to oneself could lead to better recall performance [12]. The existing literature [10] assumes that such context provided should be relevant to the incoming information. Hence, when we try to personalize learning material based on one's experience, the resultant artifact should be relevant, not just the material is personalized.

In sum, personalizing learning by providing *relevant context* based on a learner's experience is known to be effective and this is promising direction to design a effective vocabulary learning system. In this regard, utilizing Instagram posts, as already described in Introduction, is one of the enabling factor for experience-personalized vocabulary learning. To our knowledge, we could not find HCI research aiming to use such images for learning personalization, especially for vocabulary acquisition, which is one of the novelties of our work.

B. CONTEXT-AWARE LANGUAGE LEARNING SYSTEMS

As many researchers have reported, the context of a learner contains the potential to personalize learning materials. For instance, Jacquet et al. proposed *Vocabulometer*, which recommends text based on the user's vocabulary estimated from reading activity sensed by eye-tracking devices [13]. To extend their study, Yamaguchi et al. proposed *Mobile Vocabulometer*, where users choose topics of interest within the application, and it recommends English articles based on the selections to discover new vocabulary within the context [14]. The application provided the following topics: "entertainment, economy, environment, lifestyle, politics, sport, and science." While they provide preliminary work to adapt interests by user topic selection, our interest is more specific. Take the topic *sport*, for example; some people are into baseball, while others might like football. Our interests and experiences in daily life are diverse; we need to sense such backgrounds, which is another motivation for our study.

Hautasaari et al. proposed an application called *VocaBura*, which allows language learners to acquire vocabulary using their idle hours [15]. According to the study, this audio-based application allows users to find new vocabulary based on the user's location history. While they focus on personalization using location history, we focus on personalization using social image data, as discussed in the Introduction, which is another differentiator of our work.

In a similar approach to our text-generation, Arakawa et al. proposed a system, *VocabEncounter*, which aims to personalize the web-browsing experience with micro-learning design [16]. The system provides translated sentences on the learner's daily web-browsing screen. It is similar to our work in providing contextualized usages of words, but there are a few key differences. When a learner reads a web page in their native language (for example, Japanese), their system shows a sentence (or a word depending on the configuration) translated into an English sentence (or a word) containing the target English word. Their work aims to provide word usages within translated sentences, *given the target words*. In contrast, our work, Img2Vocab, focuses on finding words tied to a learner's experience using a learner's social images. Plus, while VocabEncounter focuses on technical aspects and NMT (neural machine translation) approaches to generate translations, our Img2Vocab emphasizes the creativity and opportunities of LLMs to generate content material. Hence, the ways the two systems deliver personalization, their sources of personalization, and the technical aspects of their and our interests are fundamentally different.

C. LANGUAGE MODELS AND LANGUAGE EDUCATION

The past few years have seen rapid advancements in Large Language Models (LLMs). Nevertheless, its usage for content generation or personalization for language learning, especially for vocabulary acquisition, is a newly emerged topic where we find room for exploration. In this subsection, we review recent advancements in LLMs and then discuss their current usage in content generation for language learning, illustrating the importance of our work.

Competition and advancements in research in LLMs have increased after Generative Pre-trained Transformer 3 (GPT-3) was publicly released. It is a large language model learned with 175 billion parameters [8], which has been available as APIs since 2021. With the large parameters and training data, GPT-3 is often capable of generating natural text generation, although it has room for improvements in multiple aspects, including "hallucination," where a language model generates nonsensical or unfaithful output given an input [17]. In the same year and following years, multiple LLMs with larger model sizes were introduced, thanks to Transformer architectures, which allow large input with parallelization, unlike conventional models with recurrent layers [18]. The trend also came from "scaling law" for LLMs, where Kaplan et al. suggests that the larger a language model is, the better it performs [19]. Even in 2022, some models accommodate multiple hundreds billion parameters [20], [21].

While LLMs research has been popular in NLP and other research communities, research about AI-generated material for education, particularly language learning is not thoroughly explored. From the application aspects, we can find use cases of LLMs in this domain but not all the technical details are revealed. For example, Duolingo,[3] an online language learning platform, uses GPT models (GPT-3/4) for providing roleplay practice, where a learner try to have conversation in the language that they want to learn.[4] Furthermore, GPT models have been used for Duolingo English Test according to a Duolingo Research's official report [22] and its blog post.[5] However, there are a limited number of research articles which combine LLMs and education in the language learning domain. For example, Leong et al. [23] have investigated the context generation with LLMs for vocabulary learning, and reported the positive effects of such context generation for improved learning motivation while the interventions did not see significant improvements on learners' performances, when compared to control groups. Their work is similar in that their application utilizes LLMs to generate context. However, there are two key differences. (i) They asked their participants to declare unknown words prior to the learning sessions. Their focus is to personalize context-generation given unknown words for their participants. In contrast, our research work focuses on letting a learner explore their unknown words by presenting their Instagram images and a generated text using LLMs and image recognition. (ii) Leong et al.'s [23] work focuses on text-only modality. Whereas our work utilizes both text and visual modality to enrich a learner's personalized vocabulary learning with reference to their daily lives. To our best knowledge, publicly available research work on vocabulary learning personalization using LLMs is limited since LLMs are newly emerging technology. However, we need to investigate the realm more including its effectiveness, bottlenecks for technology adoption and people's perception.

III. PROPOSED SYSTEM

This paper presents *Img2Vocab*: a first-of-its-kind system displaying a learner's images and generated sentences. As in Fig. 1, the system imports a learner's images from Instagram (left), where we can observe a learner's experience in daily life. With these images, the system generates sentences related to them, providing personalized material where learners can actively find new vocabulary that approximates their life and interest (middle). The system displays its translation by clicking an unknown word and registers the unknown word, finally providing a list of words tied to their experience (right). In our daily lives, the system plays a role in suggesting words to learn. This chapter discusses the system's overview, how it provides personalized content, and how a learner can extract words using the interface.

A. SYSTEM OVERVIEW

Our system, Img2Vocab, consists of two key components: (i) content generation and (ii) unknown word extraction. (i) Fig. 2 summarizes the content-generation pipeline, which is responsible for generating personalized content as described in the middle of Fig. 1. Another component, (ii) unknown word extraction, is also described in Fig. 3 (right). From the personalized content from the pipeline, the interface-hosted cloud allows users to interact with the content. This section discusses important steps within the pipeline and interface design.

B. CONTENT-GENERATION PIPELINE

As described in Fig. 2, the pipeline consists of multiple machine learning models on the cloud available as APIs. We summarized these models in Table 1. The pipeline generates labels from Instagram images, labels to sentences, analyzes sentences, and prepares translations.

1) DETECTING KEYWORD-OBJECTS FROM INSTAGRAM IMAGE

Given images from a learner's Instagram data, we feed them into a multi-label detection API – Google Cloud Vision API, providing up to ten labels per image with the current configuration. Take an example image from Fig. 1; the following labels were detected from the image: *Building, Window, Sky, Tree, Architecture, Neighbourhood, Travel, Public space, City, Facade*. We feed these labels into GPT-3 so that people can discover new words – (a) co-occurrences suggested by GPT-3 and (b) detected labels – related to the context strongly tied with the learner's experiences. In this example, GPT-3 provided the following sentence: "The building's facade is an intricate architecture that adds to the beauty of the cityscape."

Our content-generation pipeline does *not* perform image captioning directly. Instead, the pipeline uses object detection and passing labels into GPT-3 as a prompt. Creating a text caption from an image is already introduced, such as [24], which combines a vision CNN and a language-generating RNN. These approaches intend to generate an accurate description of an image. However, due to its design, image captioning depicts images with basic vocabulary, which might be accurate but boring.

[3] https://www.duolingo.com/
[4] https://blog.duolingo.com/duolingo-max/
[5] https://blog.duolingo.com/ai-improves-education/

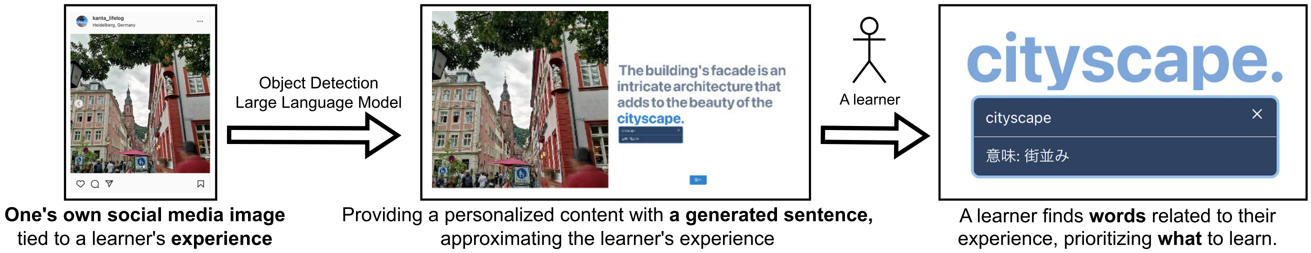

Object Detection
Large Language Model
The building's facade is an intricate architecture that adds to the beauty of the cityscape.
A learner
cityscape.
cityscape
意味: 街並み
One's own social media image tied to a learner's experience
Providing a personalized content with a generated sentence, approximating the learner's experience
A learner finds words related to their experience, prioritizing what to learn.

FIGURE 1. The proposed method where learners can actively find new words within the context of their own Instagram posts and sentences generated from GPT-3.

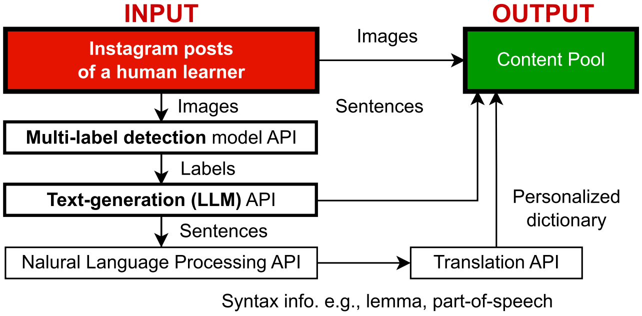

INPUT
OUTPUT
Instagram posts of a human learner
Images
Content Pool
Images
Sentences
Multi-label detection model API
Labels
Text-generation (LLM) API
Personalized dictionary
Sentences
Nalural Language Processing API
Translation API
Syntax info. e.g., lemma, part-of-speech

FIGURE 2. Overview of content-generation pipeline with multiple ML models, including multi-label detection and text generation by LLM.

On the other hand, by connecting a Large Language Model, such as GPT-3, we can add creativity to the generated text. The creativity of LLMs is mainly due to a large model and data fed into them. It is also possible to increase the randomness of a generated text by changing a parameter, *temperature*. By this design, we aim to allow learners to encounter words tied to their lives yet still new to them in a generated text with an LLM, mitigating the "exploration and exploitation tradeoff" for human language learning.

TABLE 1. External ML models or services integrated into our system.

Purpose	External Models/Services
Multi-label detection	Google Cloud Vision API
Text generation	GPT-3 Completion API (model: text-davinci-002)
Syntax analysis	Google Cloud Natural Language API
Translation	Google Cloud Translation API

2) GENERATING SENTENCES BY KEYWORDS AND GPT-3

We integrated GPT-3's completion API, which provides sentences to our system with our prompt and keyword labels. Currently, GPT-3 [8] is publicly accessible as a beta version. By nature of the text-to-text model, instructions to the model are given by sentences. Our prompt is described in Fig. 4. Lines 1-3 clarify what the model is expected to do. Our system inserts keyword labels to the line starting from *Keywords*. GPT-3 appends a sentence after *One-Sentence*. Some work, e.g., [25], tries to find a better prompt. However, we keep our prompt simple as it is not our scope of work. We used the model "text-davinci-002" from GPT-3 models as it was the latest model when we started our experiments. In Appendix, Table 4 provides the main parameters of the model. We did not finetune the language model but simply utilized the model via API. This is because we did not have an affordable open-source models which we can apply finetuning options, such as Low-Rank Adaptation [26] (LoRa), a cost effective finetuning technique, when we started the project. However, it is one of the future directions for us to use LoRa along with open sourced language models in the future, instead of OpenAI's closed models.

3) RATIONALE FOR LANGUAGE MODEL CHOICE

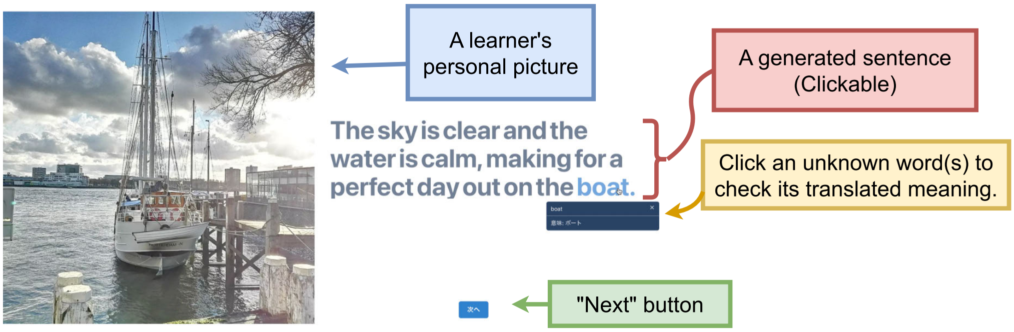
The previous setcions have clarified that we used used GPT-3 (text-davinci-002) for text-generation for our experiments. This model choice was made in early 2022, and please note that main reason for this selection was model availability in early 2022. When we started using GPT-3, there was no today's chatbot interface products, such as ChatGPT and Gemini. At that time, text-generation models generally focused on the next word prediction based on the current text input, or "prompt." As of today, however, the landscape has changed. For example, the model we used, text-davinci-002 was already deprecated and we can no longer use it for production purposes today.[6] Instead, we see new models in the form of closed source, or publically available API, or open source/publically available weights, such as GPT-4 (closed), Gemini (closed), LLaMA [27] (open), Gemma [28] (open).

That being said, there is a rationale for using simple models, such as text-davinci-002, as a hindsight. First, our system requires the text-generation model as a component of the content-generation pipeline, and we do not need models to be fine-tuned for conversations. On the other hand, we also saw, text-davinci-002, an OpenAI API being deprecated from the public API, switching to those open source models is one of our future directions for better reproducibility, but not the scope of our current work. Choosing a simple text-generation model whose weights are publically available could be a better model choice.

4) PROVIDING TRANSLATIONS FOR EACH WORD

After all the sentences are generated, we analyze syntax by using Google Cloud Natural Language API. Through this process, the API offers morphological analysis capabilities,

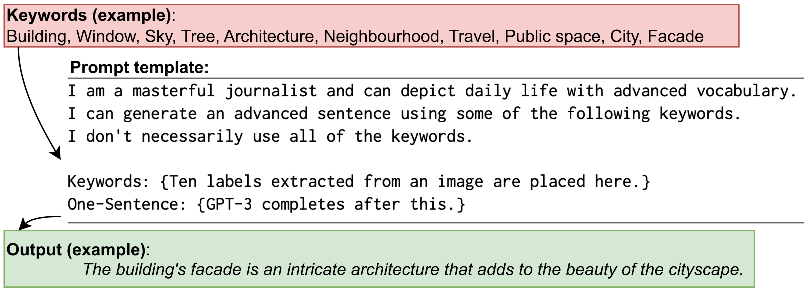
[6] https://openai.com/index/gpt-4-api-general-availability/


The sky is clear and the water is calm, making for a perfect day out on the boat.

boat ×
意味: ボート

A learner's personal picture
A generated sentence (Clickable)
Click an unknown word(s) to check its translated meaning.
次へ
"Next" button

FIGURE 3. **Word extraction interface: learners discover new vocabulary tied to their lives based on their Instagram posts and generated sentences. By clicking a word in the sentence, the word is registered as an unknown word, and the learner can check its meaning using its dictionary feature.**


Keywords (example):
Building, Window, Sky, Tree, Architecture, Neighbourhood, Travel, Public space, City, Facade

Prompt template:

```
I am a masterful journalist and can depict daily life with advanced vocabulary.
I can generate an advanced sentence using some of the following keywords.
I don't necessarily use all of the keywords.

Keywords: {Ten labels extracted from an image are placed here.}
One-Sentence: {GPT-3 completes after this.}
```

Output (example):
The building's facade is an intricate architecture that adds to the beauty of the cityscape.

FIGURE 4. **Our prompt template inserts keywords extracted from an image. GPT-3 completes after the last line, providing a creative sentence related to keywords.**

providing us with the following helpful information for our system: part-of-speech and lemma. We need lemmas to prevent the system from extracting multiple but essentially the same words, such as *car* and *cars*. Users can click each sentence word within our system to learn them with standard forms and the corresponding translation of the standard forms. Google Cloud Translation API provided translations for these standard forms.

5) UNKNOWN WORD EXTRACTION

Our interface to extract unknown words is described in Fig.3. In the proposed setting, it displays a learner's image on the left and a generated text on the right. Learners are supposed to find unknown words in the generated texts related to daily life. Once they click a word in a sentence, the system registers it as an unknown word. At the same time, the system displays a translation of a word. By this design, we expect a learner *extracts* new words to learn that are tied to their daily lives. This is the main aspect of "unknown word extraction." Note that our primary focus is to allow learners to *find* new words tied to their experience as a first step. It is rather agnostic to specific ways to learn by heart; once we *extract* words to learn, they are their own choice, e.g., using flashcards or spaced repetition.

IV. OVERVIEW: PILOT AND LARGE-SCALE STUDY

We carried out two studies, a pilot study and a large-scale study, to clarify the effectiveness of our system – as a tool to extract words tied to our experience. Two experiments can be summarized as follows:

E1 Our first pilot study with three participants
E2 Our large-scale experiment conducted with 23 participants

With the pilot study (E1), we aimed to determine the potential bottleneck in data collection for our experiment with the initial prototype. The pilot study also gave us insights into the experiment design, such as the scalability or effectiveness of the data collection procedure. Based on these lessons, we refined the design and conducted another large-scale experiment (E2) with more participants.

In this paper, we mainly explain E2, the large-scale study, because E2 has more participants and improvements in the design, which we explain in the next section. It is noted that the initial report of E1 and results were elaborated on our previous work [29]. This paper is an extension of the previous work. However, most of the extension work is based on a large-scale study rather than the initial pilot study.

A. THREE CONDITIONS FOR OUR EXPERIMENTS

To clarify how our proposed personalization can affect the exploration of unknown words – compared to otherwise (without personalization), we prepared three conditions, C1-C3, during the two studies. The three conditions for our experiment are summarized in Fig. 5 and as follows:

C1 Displaying an image provided by the user and a sentence generated from the image (Proposed).

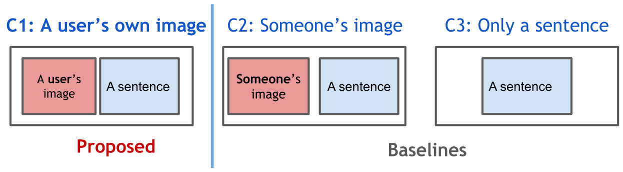


FIGURE 5. Three conditions we prepared for our experiments: C1-C3.

C2 Displaying an image provided by someone and a sentence generated from the image (Baseline 1).
C3 Displaying only a sentence generated by an image provided by someone (Baseline 2).

Our proposed interface, described in Fig. 3, was based on the primary hypothesis, "Displaying a picture taken by the user and related to a sentence is useful in exploring unknown words for language learning." Two essential factors to consider are: (1) Displaying one's image or someone else's image along with the text. (2) Displaying images along with text or only text. With this in mind, we added two baselines during the studies, C2 and C3. After the experiments, we compared values gained, such as difficulties of words, over C1-C3.

Content from C3 derives only from someone's image, but the image is invisible. We did not add another possible condition with "only a sentence generated by an image provided by the user" due to the limited number of images provided by learners. In the pilot and the large-scale study, we required every participant to have at least 30 images to join the studies. On the other hand, the authors provided pictures for C2 and C3 during our study.

B. THE MAIN TASKS FOR OUR EXPERIMENT

Our two studies have the same steps in common in the flow to evaluate our primary hypothesis: "Displaying a picture taken by the user and related to a sentence is useful in exploring unknown words for language learning." Fig. 6 summarizes these common steps, which we call "the main tasks." Mainly, from the main tasks, we aimed to know how our proposed condition C1 differs from other conditions regarding types of extracted words and memory retention after immediate distraction tasks.

Participants clicked unknown words in the learning phase to check their meanings, as in Fig. 6. After selecting all unknown words in one sentence, participants clicked the button to display the next sentence. This process continued until the number of unknown words reached 10 in three conditions (30 words in total) or all images prepared for C1 were shown. As in the middle of Fig. 6, the system displayed a distraction task after users extracted unknown words from the three conditions. The simple addition task of three one-digit numbers was carried out for over 1 minute. The following recall task prompted the user to spell out English words within the user's unknown word list in a shuffled order, providing scores for each condition to capture memory retention.

V. LARGE-SCALE STUDY AND RESULTS

A. DESIGN OF LARGE-SCALE STUDY

We describe the large-scale study (E2) by illustrating the essential similarities and differences between the two experiments, as we already discussed the design of the pilot study (E1). When comparing E1 and E2, we can summarize the key differences: participants, data collection improvement, experimental flow, and added questionnaires. This large-scale study followed the ethics protocol within DFKI with the consent of the participants.

1) PREREQUISITES AND COMPENSATION FOR PARTICIPANTS

For E2, we recruited 23 Japanese participants living in Japan, not limited to university students, diversifying our experimental demographics. The average age for participants in E2 was 26.8 (STD=4.49). Participants consist of 9 females and 14 males. We listed further details of participants in E2 to Appendix. (See Table 8.) Note that participants in E1 and E2 are different groups, although we describe participants in E2 with the same notation as our description of E1 – "P1, P2, P3, ..."

Another new requirement for E2 is to accept permissions to their Instagram posts via its API to automate image collection from Instagram.[7] Also, participation in E2 was voluntary without compensation.

2) IMPROVED DATA COLLECTION PROCEDURE ON THE CLOUD

Integration to Cloud Authentication and Backend – The system in E2 is more production-ready using a public authentication provider — as shown on the left in Fig. 7. With this change, the system could securely store data, such as exported images and extracted words. This improvement reduced manual operations by both participants and authors drastically in E2. It can automatically start the content-generation process and guide participants to the main task once the generation has finished for each person. Also, the system automatically collects data, e.g., extracted words, on the cloud instead of asking them to send us the exported CSVs from the system.

Summary: Experimental Flow Overview – We summarized the improvement we discussed so far in Fig. 8. In Fig. 8, the interface improvements we discussed are highlighted in blue. The main task we discussed in the pilot study in Fig. 6 corresponds to the green steps in Fig. 8. We also highlighted newly added questions in yellow, which we call *extra-questionnaires* and which we address later.

3) ETHICAL CONSIDERATIONS

We discuss our ethical considerations we took for our system to carry out experiments with intention to promote broader adoption of technology. Our system utilizes highly personal information, images from a learner's Instagram. This entails

[7] https://developers.facebook.com/docs/instagram-basic-display-api

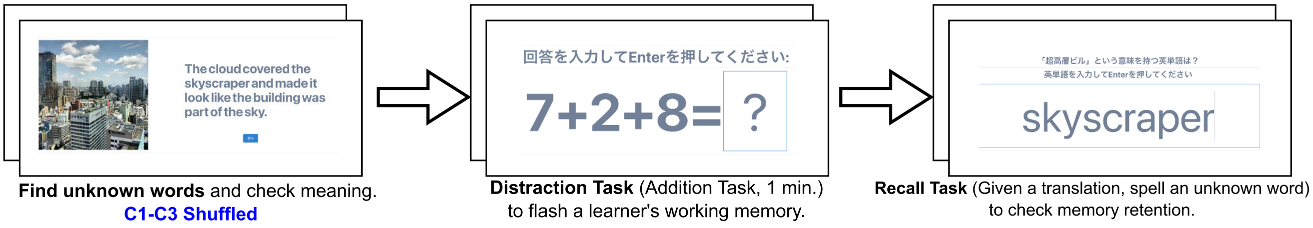
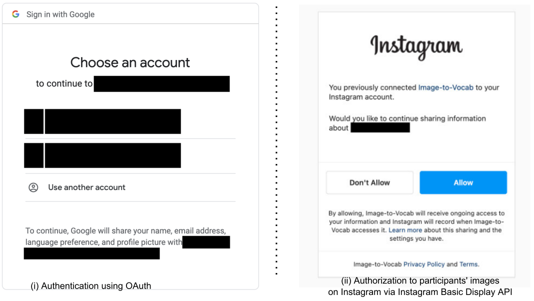

The cloud covered the skyscraper and made it look like the building was part of the sky.
回答を入力してEnterを押してください:
7+2+8= ?
「超高層ビル」という意味を持つ英単語は？
英単語を入力してEnterを押してください
skyscraper
Find unknown words and check meaning.
C1-C3 Shuffled
Distraction Task (Addition Task, 1 min.)
to flash a learner's working memory.
Recall Task (Given a translation, spell an unknown word)
to check memory retention.

FIGURE 6. Overview of the main part of the evaluation procedure for the pilot study. (Also the same for the large-scale experiment.)


Sign in with Google

Choose an account

to continue to [redacted]

[redacted]

Use another account

To continue, Google will share your name, email address, language preference, and profile picture with [redacted]

(i) Authentication using OAuth

Instagram

You previously connected Image-to-Vocab to your Instagram account.

Would you like to continue sharing information about [redacted]

Don't Allow | Allow

By allowing, Image-to-Vocab will receive ongoing access to your information and Instagram will record when Image-to-Vocab accesses it. Learn more about this sharing and the settings you have.

Image-to-Vocab Privacy Policy and Terms.

(ii) Authorization to participants' images on Instagram via Instagram Basic Display API

FIGURE 7. Integration of the third-party authentication provider for a seamless and secure experiment (left) and Instagram API for seamless import of images from a learner's Instagram (right).

the inherent nature of personalization and privacy tradeoff. One of the existing privacy frameworks [30] emphasizes the importance of providing user control of their data for the trustworthiness of the system. In this regard, our system has an interface to allow users to opt-in images for better user control of their privacy. Also, please note that the data is only used for the learner's personalization, which is classified to have more user control according to the privacy framework [30].

User Interface to Opt-in Images – Unlike the prototype system in E1, the new system in E2 can import Images using Instagram Basic Display API as in Fig. 7 (right). In addition, we added an interface to opt-in images that participants want to share via our system as in Fig. 9. Since Instagram posts can sometimes be private, allowing users to choose what they disclose is essential from a privacy perspective.

Data Privacy and Security Measures— Since images from a user's social media can be very personal, we carefully designed our system with data privacy and security. The images we obtained via Instagram Basic Display API were stored in Cloud Storage. All the communications were encrypted via HTTPS. Our database or storage were securely protected with a security policy tied to the authentication provider, which allowed access to user-specific data to that particular user, not to the rest of the user pool. The vast majority of processes were done within Google Cloud/Firebase. For text-generation, we used OpenAI API. Those cloud platforms have fulfilled compliance standards and certificates.[8,9] We combined those components with with care (e.g., with correct configuration and security policies), hence our system robustly protects private data.

4) EXTRA-QUESTIONNAIRES

In the large-scale study, we added several questionnaires to users before and after the main tasks. We highlighted these three additional studies with yellow in Fig. 8, and we discuss each in this subsection.

Pre-questionnaire (Demographic) – Before the main tasks, we asked participants a demographic questionnaire via an online form written in Japanese. It was mainly about participants' English proficiencies and how frequently they study English. We listed all the items, except an item to collect contact information, in the Appendix; see Tabel 5.

User Study (User-dependent Content) – In our software, we added a user study after the recall task at the end of our experiment. The purpose of our user study is to clarify the following using participants' subjective reports:

1) The usefulness of each extracted word within sentences during the learning phase.
2) A user's interest in each image they had seen during the learning phase.

First, the metric 1 is a vital criterion for evaluating our system. This is because our system intended to let users discover vocabulary related to their daily life. Via the metric 2, we wanted to know how much our proposed condition C1 captured learners' interest by personalization with their images.

We embedded 1-5 Likert scale items in the system to clarify the two criteria. We listed actual items in the Appendix. The system displayed the items after the recall task but not during the task. We placed items in ascending order in our Likert scale to avoid "left-side selection bias," indicated by previous research such as [31]. Concretely, selections "1," "3," and "5" corresponded to strong disagreement, neutral opinion, and strong agreement. For metric 1, the system asked users the question (See Q_a1 in Table 6) for every extracted word during the user's experiment. For metric 2, it asked them corresponding questions (See Q_a2) for every image a user had seen during the experiment – corresponding to C1 and C2.

Post-questionnaire – In addition to the questions in the previous subsection, we asked every participant to answer an online form after the experiment. Table 7 in the Appendix describes the content of our post-questionnaire. It contains both Likert items and free writing items. Likert items had the same rules as we discussed in the last paragraph. The

[8] https://cloud.google.com/compliance

[9] https://trust.openai.com/?itemName=certifications

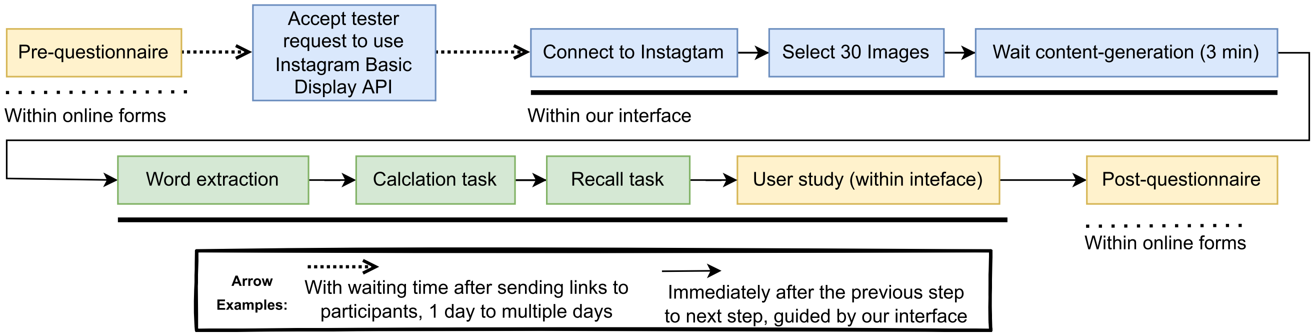
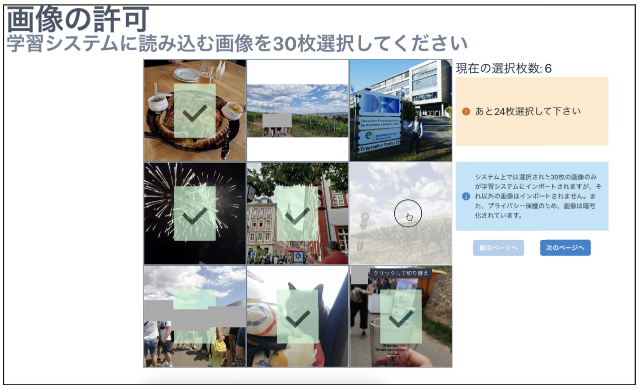

Pre-questionnaire
Within online forms
Accept tester request to use Instagram Basic Display API
Connect to Instagtam
Select 30 Images
Wait content-generation (3 min)
Within our interface
Word extraction
Calclation task
Recall task
User study (within inteface)
Post-questionnaire
Within online forms
Arrow Examples:
With waiting time after sending links to participants, 1 day to multiple days
Immediately after the previous step to next step, guided by our interface

FIGURE 8. Experimental timeline for our large-scale experiment.


画像の許可
学習システムに読み込む画像を30枚選択してください

現在の選択枚数: 6

あと24枚選択して下さい

システム上では選択された30枚の画像のみが学習システムにインポートされますが、それ以外の画像はインポートされません。また、プライバシー保護のため、画像は暗号化されています。

FIGURE 9. Our interface allowed learners to opt-in images they wanted to share via our system during the large-scale study. The system imported only 30 opted-in pictures for the entire session to respect a learner's choice and privacy.

questionnaire was mainly about: the helpfulness of conditions C1-C3, privacy concerns about using images on their social media, how much they enjoyed the proposed methods, and other miscellaneous questions.

B. RESULTS OF LARGE-SCALE STUDY

1) UNKNOWN WORDS EXTRACTED BY PARTICIPANTS

Exclusion and Sample Size – The number of participants who completed the large-scale study was 23, and we gained a sample size of $n = 21$ after excluding two. We discuss the reason for this exclusion and its implication. To compare three conditions C1-C3, we needed words in every condition in each participant. In the experiment, however, two participants could not extract words for every condition. Thus they were excluded from our further score analysis. These two participants learned all the pictures in C1, which led to an early exit from our experiment before collecting at least a word for every class. One of the interpretations for this is that their English proficiency was possibly too good to get a sufficient number of unknown words. Although we excluded two from the analysis, we still used their data from user studies.

The Number of Extracted Words – The word count for extracted words per class is described in Fig. 10. The average word count for C1 is slightly lower than for other classes. Participants collected fewer words from such context than in other conditions. They could be already familiar with related words in a context similar to their daily lives (C1). Nevertheless, regarding the standard error of the mean indicated by the red bars in the Fig.10, statistical power is limited, we note.

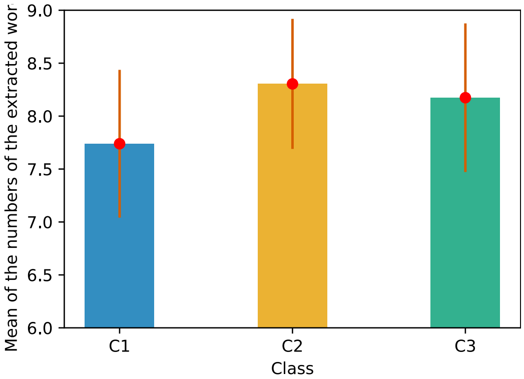


FIGURE 10. Extracted word count for each participant with standard error of the mean (SEM) error bars in the large-scale study (n=23).

The Difficulty of Extracted Words From the Corpus – We calculated Zipf values for the 21 participants as shown in Fig. 11 to see the difficulties of extracted words. We used Zipf frequency, which we calculated by taking the logarithm, with base 10, of the number of times a word appears per billion words. In some words, the Zipf values could be zero—not match in the corpus. In such cases, we removed the word from the mean calculation of Zipf values. We used a python package called "wordfreq" [32] to calculate the Zipf values of extracted words. Lower Zipf values indicate words are less likely to appear, which is considered more difficult. In Fig. 11, the mean Zipf value for C3 (only text) is highest among C1-C3. In C3, participants tended to mark easier words as unknown words. One of the explanations for this is that C1-C2 had images along with the texts, which allowed them to guess their meanings based on the image context. Therefore, materials with images (C1-C2) prioritized words that are difficult to guess the meaning in their daily lives. The word frequency comparisons support C1-C2, meaning having images allow learners to find more difficult vocabulary than having text-only context.

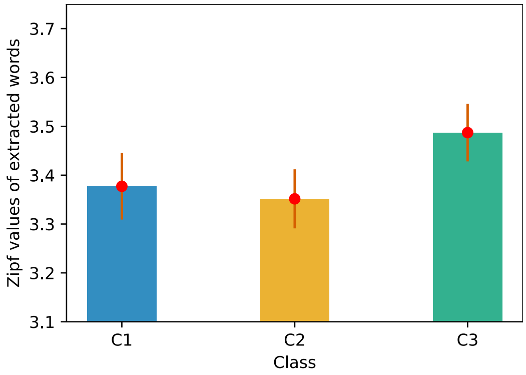


FIGURE 11. Frequencies of the extracted words over three conditions with standard error of the mean (SEM) error bars: We used Zipf frequency to compare word difficulty over three conditions. (21 participants).

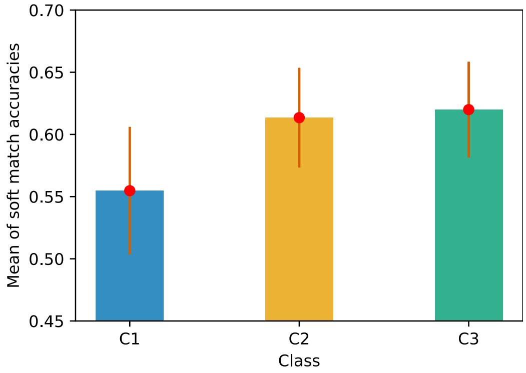


FIGURE 12. Accuracies (soft match: considering misspells) over three conditions with standard error of the mean (SEM) error bars (21 participants): Measure for memory retention after immediate distraction tasks.

2) MEMORY RETENTION AFTER IMMEDIATE DISTRACTION TASKS

In this subsection, we first explain how we calculate accuracies while considering misspells – which we refer to as "soft match accuracy," and then we compare such accuracies over three conditions.

Calculation of Accuracies: Full Match vs Soft Match – One of the ways to know how accurately the participant spelled out is a strict checking where we divided the number of correct answers by the number of extracted words within the same class – which we refer to as "full match accuracies." We used this measure for the pilot study. However, it is essential to know how far almost a user got the answer correct. Take the word "coffee," for example; the misspelled word "cofi" should be penalized more than another misspelled one, "cofee." To mitigate this, we mainly compare the soft match accuracies in our large-scale study, which we explain in the next paragraph. The soft match accuracies can consider participants' output in a detailed gradation rather than the full match accuracies.

The Rules for Soft Match Accuracy Calculation – We converted the user's spellings and expected answers into lowercase before comparing a pair, gaining s_1 and s_2. Instead of strict comparison as in full match accuracy, we considered misspells for each condition with the Levenstein distance l in soft match comparison. To get soft match accuracies of a condition per user: we calculated a function $f(s_1, s_2)$ as in (1) where $|s|$ indicates the length of a word s, for every pair of a user's spelling s_1 and an answer s_2 (both lowercase), accumulated them over conditions and finally divided by the extracted word count per user.

$$f(s_1, s_2) = 1 - \frac{l(s_1, s_2)}{\max(|s_1|, |s_2|)} \tag{1}$$

The fraction on the right side is also known as the normalized Levenstein distance. The range of $f(s_1, s_2)$ is $[0, 1]$. When the two strings are the same, this gives 1, whereas if they are different, the function $f(s_1, s_2)$ provides $[0, 1)$ values depending on how different s_1 and s_2 are. Soft match accuracies are always equal to or greater than the full match accuracies.

Comparisons of Accuracies – We calculated three classes' average soft match accuracies as in Fig. 12. Among average accuracies C1-C3, the condition for the proposed system, C1, yielded lower accuracies than the rest; Having an image that is also one's own, users could extract difficult words that they could not easily learn by heart. To illustrate the statistical difference in soft match accuracies among the three classes, we conducted one-way ANOVA, which we summarized into Table 2.We used the statsmodels[10] packages to carry out one-way ANOVA with the null hypothesis that "two or three groups have the same population mean."After checking the prerequisites for the test, namely homoscedasticity and normality, one-way ANOVA for the soft match accuracies yielded $F = 0.6755$ and $p = 0.5127$. With significance levels $\alpha = 0.05$, this can not reject the null hypothesis. Therefore, in our study, C1 yielded lower accuracies than C2 and C3 with insignificance. In other words, having a context grounded with a learner's social image allowed them to find difficult words to quickly learn than having context generated by someone's image, or text only modality—although this finding is statistically insignificant.

TABLE 2. The results of ANOVA for soft match accuracies among three conditions, C1-C3. SS and DF stand for sum of squares and degrees of freedom respectively.

	SS	DF	F	P-Value
Factor	0.054227	2.0	0.675483	0.512738
Residual	2.408386	60.0	-	-

3) EXTRA-QUESTIONNAIRES

Unlike results from the main tasks, where we excluded two participants, we included responses from all participants, hereafter providing a user study from 23 participants with the questionnaires. In our questionnaires, some items are

[10] https://www.statsmodels.org/stable/generated/statsmodels.stats.anova.anova_lm.html

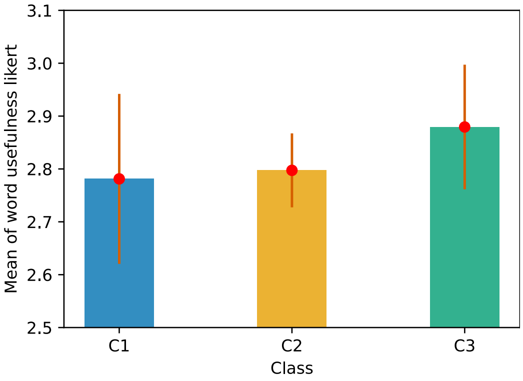


FIGURE 13. The usefulness of words reported by users from 1-5 Likert scale with standard error of the mean (SEM) error bars over C1-C3.

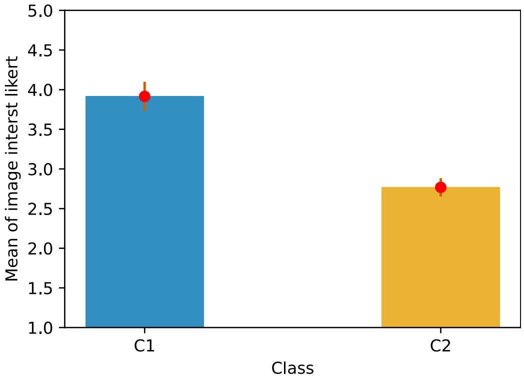


FIGURE 14. The interest of images reported by users from 1-5 Likert scale with standard error of the mean (SEM) error bars over C1-C2.

1-5 Likert scale. On the scale, strong disagreement with the statement given corresponds to 1, a neutral opinion corresponds to 3, and strong agreement corresponds to 5.

User Study Within System (User-dependent Content) – First, we calculated the average of the Likert item (see item Q_a1 in Table 6 of Appendix), which asked how useful each extracted word is in learners' daily lives, along with translation. We first calculated average Likert values for each person and then calculated the average on each condition. We visualized the result in Fig. 13. Note that each participant has extracted different words; therefore, sets of extracted words for P1 and P2 are different; for example – the items were user-dependent.

There are mainly two findings. (i) We observed that the standard error in C1 is higher than in C2 and C3. It is difficult to interpret this result. However, people who felt useful and not so useful were mixed. The reason for this remains unclear. (ii) average values are the biggest in C3, followed by C2 and C1. People may say easier words are useful because the shape of Fig. 13 is similar to Zipf values in Fig. 11. This might be because they can imagine their usefulness easily for easier words. This result indicates we have to design questions carefully to know about the *usefulness* of words in people's daily lives, as people are likely to say easier words are *useful*, which is not necessarily the case.

Next, we calculated the average of the Likert item for every image the system displayed during the study. The item (see item Q_a2 in Table 6 of Appendix) asked about participants' interest in each image. We visualized the result in Fig. 14. The average calculation was done similarly to the word usefulness Likert. From Fig. 14, we observed that C1 yielded a higher average than C2, indicating a learner's own Instagram images capture their interest compared to someone's images. This supports our idea of using personal pictures as a data point for learning personalization, as we expected.

Post-Questionnaire – For Likert scale items in Table 7, we calculated the average of each participant's response and analyzed the values assigned to each scale, 1-5. In our post-questionnaire, we listed only Likert items and their average values in Table 3.

TABLE 3. Average 1-5 Likert scale values from our post-questionnaire.

Name	Question (Originally written in Japanese)	Average
Q_b1 (C3)	Regarding the times when text-only example sentences were displayed when learning vocabulary in this application, do you think such content helped you learn? (1: least, 5: most)	2.826
Q_b2 (C2)	Regarding when images of other people's Instagram posts were displayed along with example sentences in learning vocabulary in this application, do you think such content helped you learn? (1: least, 5: most)	3.522
Q_b3 (C1)	Regarding when your image was displayed with example sentences in learning vocabulary in this application, do you think such content helped you learn? (1: least, 5: most)	4.304
Q_b4	Do you have any privacy concerns about using your Instagram as your own personal learning resource? (1: least, 5: most)	2.435
Q_b7	Did you enjoy learning vocabulary in this application by using your own Instagram posts as your own personal study material? (1: least, 5: most)	4.043
Q_b9	Do you think the application used in this experiment was easy to use? (1: least, 5: most)	3.652

First, we explain the three items (Q_b1-Q_b3) and their implications. We designed these items to compare which conditions were helpful to learners using subjective reports. Concretely, the items Q_b1, Q_b2, and Q_b3 correspond to C3, C2, and C1. By this design, the closer to 5 the average is, the more helpful the corresponding condition is to learn vocabulary as participants' subjective report. The averages are as follows: C1 (proposed): 4.304, C2: 3.522, and C3: 2.826, indicating our proposed method C1 is the most useful among others from the subjective report. From this, C1 (having an image along with text) was most helpful, and C2 (having someone's image and text) followed. Condition C1, only text, yielded the lowest average.

Second, we explain the item regarding privacy concerns Q_b4. We added this item because we also wanted to know people's acceptance of using a learner's Instagram post as input to our system. The average value of Q_b4 is slightly below 3, indicating participants are less likely to mind such data collection for enhancing their language abilities.

However, we cannot generalize this result to people outside the participants, as it is also possible that only those with less privacy concern were likely to participate in our experiment, which could be a sort of "survivorship bias." Therefore, designing the architecture with data security in mind is still important. We observed some privacy concerns that participants or participant candidates noted. For example, a participant noted in the free writing section Q_b5, corresponding to Q_b4, "For my public posts, I do not care about privacy at all. However, I might be concerned about privacy for my private posts." In addition, when we tried to recruit participants, some people were concerned about data privacy even after we explained the data collection, purpose, and data protection strategy. We did not count how many refused to join our participants due to privacy concerns. Personalization plays a vital role in enhancing the learning process. However, at the same time, clarifying transparency about how the system and developer protect data is equally important as making the system itself secure – to make such a system acceptable to people.

Another Likert item, Q_b7, asked if participants enjoyed learning vocabulary in the proposed setting (using your own Instagram as material); the average was 4.043. Also, participants provided an average of 3.652 to the item Q_b9, which asked about the application's usability in the experiment. Participants mainly reported positive impressions toward our proposed system from Likert items in our post-questionnaire. Although participants' feedback is somewhat biased, this psychologically positive feedback might affect the long term, so it might be helpful to experiment in the long term. For example, experimenting with the between-group design: one group of people with C1 and the other with C2 in the long term, e.g., one month.

C. ROOM FOR IMPROVEMENT OF THE SYSTEM

1) LACK OF WORD TRANSLATION BASED ON THE SENTENCE CONTEXT

When we translate one language to another, some vocabulary, such as proper names or technical terms, does not exist in the other language. Because of this, we often use phonetic translation, especially between languages with different sound systems—such as English to Japanese, and this translation is specifically called "transliteration" [33].

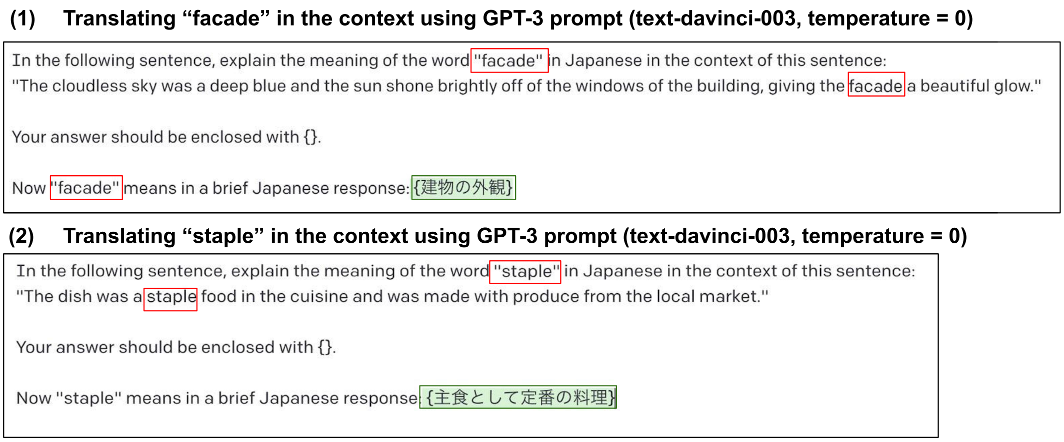
However, our system faced some unwanted transliterations due to translation design; Our system provided translations of each clicked word by passing to Google Cloud Translation API. However, this translation in the pipeline needed to have considered the meaning in the context of the generated sentence ideally. We observed some unwanted transliterations within our system. For example, the translation API mapped "facade" (English) to "fasado" [fasãdo] (Japanese) in Japanese phonetic symbols—called "Katakana." In addition, eight out of 23 participants mentioned (approx. 35%) the translation quality, while one said it is very convenient because the person did not have to look up the meaning in the dictionary. One of them described in the user study, "I know it cannot be helped as long as machine translation is used, but I was concerned that some Japanese translations were in Katakana."

As a solution, we also found that we could mitigate this using GPT-3 with a simple prompt technique. After our experiments, we noticed that a new version of GPT-3 (text-davinci-003) can also be used to provide translation of a word in a sentence and sentence generation. Take the words "facade" or "staple" as an example; our system did not provide correct translations to them during our experiments with an external translation API. On the other hand, if we use the prompt as in Fig. 15 with GPT-3 text-davinci-003, we can observe that the GPT translates the words in the real generated sentence with the context in mind. It also understands the instruction to enclose translation with "{}," where we can programmatically extract the translation from the output of GPT. Although we need further evaluation of how accurately GPT or LLMs can translate words partially in the context, integrating them into our pipeline to provide translations is one of the promising directions.

Instead of the prompt technique of LLMs above, utilizing NLP approaches also seems promising in providing a relevant word translation within a sentence. For example, an idea would be the following: considering a word and its attention mechanism [34] within original (English) and translated sentence (Japanese) pairs, then using attention relationships between the extracted English word and find its corresponding word in the translated Japanese word. This way, we might be able to obtain the translation of a word within a sentence. To sum up, while our scope for the current paper is not finding optimal NLP techniques or prompt techniques for better translation, we can employ either to improve our system.

2) ENRICHING CONTEXT BY ADDING VISION LANGUAGE MODELS

In our current work, using the optimal LLMs or image multi-class labeling was not the scope because we focused on our initial proof-of-concept. We also kept the content generation pipeline intentionally simple, as in Fig. 2, because the life cycles of those technologies are short and rapid. As in Fig. 2, we first employ the multi-class labeling and then feed those labels to LLMs to generate texts. The newly emerging technology, Vision Language Models (VLMs), promises to replace our simple content-generation design, enriching capturing a person's life with higher resolutions. For example, the multi-label detection API may provide generic object labels such as "building" as Fig. 4. In contrast, however, VLMs are capable of capturing nuanced semantics of the objects in an image (or even videos) and also can directly generate sentences given an image and a prompt [35], [36]. Since some VLMs are publicly becoming available as



(1) Translating “facade” in the context using GPT-3 prompt (text-davinci-003, temperature = 0)

In the following sentence, explain the meaning of the word "facade" in Japanese in the context of this sentence:
"The cloudless sky was a deep blue and the sun shone brightly off of the windows of the building, giving the facade a beautiful glow."

Your answer should be enclosed with {}.

Now "facade" means in a brief Japanese response: {建物の外観}

(2) Translating “staple” in the context using GPT-3 prompt (text-davinci-003, temperature = 0)

In the following sentence, explain the meaning of the word "staple" in Japanese in the context of this sentence:
"The dish was a staple food in the cuisine and was made with produce from the local market."

Your answer should be enclosed with {}.

Now "staple" means in a brief Japanese response {主食として定番の料理}

FIGURE 15. Translation of a word’s meaning in the context using GPT-3 prompt. (text-davinci-003, temperature = 0) These examples are real generated sentences during our experiments where we observed failure translating the word (1) “facade” or (2) “staple” due to unwanted transliteration.

TABLE 4. Overview of parameters for GPT-3 for text generation for our system.

Parameter	Value
Model	text-davinci-002
`temperature`	0.5
`max_tokens`	100
`top_p`	1
`frequency_penalty`	0
`presence_penalty`	0

APIs [37], [38], incorporating those VLMs into our content generation will enrich the learning context that the system can offer. This is one of the future directions for improving our system.

3) LIMITATION OF THE SAMPLE SIZE AND ITS IMPLICATION

Our experiment hosted 23 participants. One may say this is a relatively small sample size. We would like to clarify the possible reason for this and its implications. While our qualitative studies show positive reviews from the participants, we had difficulty collecting participants with often refusals. We attribute this to the nature of the experiment requiring one of the most sensitive information, a participant’s Instagram photos. We did not count the success and failure rate during participants recruiting, and did not ask follow up questions of potential participants who refused, so we cannot discuss in a quantitative fashion. In contrast, our participants generally mentioned the system positively after the experiment. This qualitative observation implies that not only designing a system in a privacy-preserving way matters, but also thorough explanation of how their data are used and how they can be benefited by having such a personalization prior to their decision making on whether they participate or not. We also discuss more about privacy implications next.

TABLE 5. Items in our pre-questionnaire (demographic).

Name	Item (Originally written in Japanese)
Q_d1	Your age
Q_d2	Your gender (“female”, “male”, or “N/A”)
Q_d3	Your scores of English proficiency tests, if any. (Optional, free writing)
Q_d4	Do you study English on a regular basis? (“everyday”,“several times a week”, “almost never”, “never”).

TABLE 6. Questions for user study within our system.

Name	Question (Originally written in Japanese)
Q_a1	To what extent do you think this word has a use in your daily life? (1: least, 5: most)
Q_a2	To what extent are you interested in this photo? (1: least, 5: most).

4) FURTHER ETHICAL CONSIDERATIONS FOR THE FUTURE OF LEARNING PERSONALIZATION WITH PRIVATE INFORMATION

While we made efforts to mitigate privacy and ethical concerns to design our system, which we already elaborated in the proposed system section, there are several aspects to be further improved for better technology adoption for society. Personalization inherently entails privacy risks and it is important to align with existing frameworks. To our knowledge, however, we could not find literature to examine this in the language education settings, so we had to follow a general framework for privacy in the personalization system [30], which was developed over time with existing privacy matters, such as online advertisement. This means we need a privacy framework for personalization using sensitive data to generate context for education. This future work needs

TABLE 7. Content of our post-questionnaire for the user study.

Name	Question (Originally written in Japanese)
Q_b1	Regarding the times when text-only example sentences were displayed when learning vocabulary in this application, do you think such content helped you learn? (1: least, 5: most)
Q_b2	When images of other people's Instagram posts were displayed along with example sentences in learning vocabulary in this application, do you think such content helped you learn? (1: least, 5: most)
Q_b3	Regarding when your image was displayed with example sentences in learning vocabulary in this application, do you think such content helped you learn? (1: least, 5: most)
Q_b4	Do you have any privacy concerns about using your Instagram as your own personal learning resource? (1: least, 5: most)
Q_b5	Please describe any feelings you have about using your Instagram as your own personal learning resource, if any. (Optional, free writing)
Q_b6	When do you post images to Instagram? (Optional, free writing)
Q_b7	Did you enjoy learning vocabulary in this application by using your own Instagram posts as your own personal study material? (1: least, 5: most)
Q_b8	Any reason you can give for your answer to the previous question? (Optional, free writing)
Q_b9	Do you think the application used in this experiment was easy to use? (1: least, 5: most)
Q_b10	Please let us know if there are any problems or improvements in the applications used in this experiment. (Optional, free writing)

to be done by comprehensive research communities including educators, pedagogist and human-computer-interaction researchers. There are many open-ended questions that arise, such as, do we really need such detailed personal information to personalize learning at the expense of a learner's privacy? What are the technical safeguards can we have to mitigate this? These questions need to be addressed down the road.

Aside from the need for a new privacy framework for education, we still have several aspects of our system to improve based on the existing general framework [30]. First, while our system utilized trustworthy and secure cloud computing services for quick prototyping and scalability of our experiment in mind, client-side computing is desired according to the existing privacy framework [30]. One of the existing studies of context-based personalization, (author?) [16], also mentioned this as a future direction. We would be more aligned with research communities by introducing client-side computing. Currently, our system utilizes image recognition and text-generation models from cloud providers. In the future, on-device image recognition [39] and on-device LLMs, which runs on a smartphone [40], for example, will more accessible in the future. With that, we may replace our pipelines in the cloud into a smartphone. Another aspect of our system to improve is to introduce technical safeguards for privacy. For example, the existing privacy literature [30] mentions that social-based personalization—a personalization using a person's social media data—might expose not only the person's privacy, but also their friends' privacy. In our system, if a learner selects an image with their friends, it's possible that their faces were used without consent although the material is only visible to the learner. In the future, we can include a component in the pipeline to detect sensitive information, such as people's faces, and either remove those photos from the material or add blurs to them until we can develop a feasible client-side system. The two aspects of our system improvement aforementioned will lead to better acceptance of the technology.

TABLE 8. Participants information collected for the large-scale study. "TC," "TF," "IT," and "EK" stands for "TOEIC," "TOEFL," "IELTS," and "Eiken," respectively.

Id	Age	Gender	English Proficiency	English Study
P01	22	M	TC 990, EK 1	Several times a week
P02	25	M	-	Almost never
P03	42	M	TC 870	Everyday
P04	23	M	TC 730	Everyday
P05	24	F	IT Overall 7.5	Almost never
P06	27	M	-	Everyday
P07	28	M	TF iBT 87	Everyday
P08	29	F	TC 485	Almost never
P09	29	M	-	Never
P10	29	F	TC 885	Almost never
P11	28	M	TC 780	Almost never
P12	26	M	TC 550	Several times a week
P13	24	M	TC 810	Never
P14	25	F	TC 795, TF iBT 94	Everyday
P15	28	M	TC 990, EK 1	Everyday
P16	33	M	TC 870	Several times a week
P17	28	F	EK 2	Almost never
P18	22	F	IT Overall 7	Several times a week
P19	29	M	TC 860	Several times a week
P20	28	M	TC 960	Several times a week
P21	24	F	TC 885	Almost never
P22	21	F	TF iBT 105	Almost never
P23	22	F	TC 890, IT Overall 6.5 (Academic)	Several times a week

VI. CONCLUSION

Personalization of learning based on one's experience is vital for vocabulary acquisition. This learning personalization entails two challenges: *sensing* a learner's experience and *creating* material for each individual. To mitigate this, we developed a system displaying learners' images and generated sentences from our ML pipeline, including GPT-3, grounded with a learner's social media images from their Instagram. This way, learners can actively find new vocabulary that approximates their lives and interests. We carried out our within-subject design evaluation of the system with 23 participants with three conditions. With comparisons of the recall tasks accuracy, we found that having a context generated from a learner's social image encouraged them to find more challenging words to quickly learn than having context from images that are not related to them, or text only modality, with insignificant differences reported. The Zipf frequency comparisons also revealed generally having image-based context allowed learners to explore difficult vocabulary than having text-only material. We also discuss quantitative and qualitative results and implications towards the acceptance of the personalization system with their personal photos from social media. While they reported positive feedback for our system, we also discuss future improvement aspects such as technical aspects and additional ethical considerations. These findings and implications will contribute to a future with a better personalized vocabulary learning, tied to a learner's real-world experiences.

APPENDIX A
TEXT-GENERATION WITH GPT-3

See Table 4.

APPENDIX B
CONTENT OF QUESTIONNAIRE

See Tables 5–7.

APPENDIX C
DEMOGRAPHIC INFORMATION

See Table 8.

ACKNOWLEDGMENT

The authors would like to thank Andrew W. Vargo and Javier Carrasco Melo for tips on within-subject statistical analysis and recommendations for Python statistics packages.

REFERENCES

[1] G. Brod, M. Werkle-Bergner, and Y. L. Shing, "The influence of prior knowledge on memory: A developmental cognitive neuroscience perspective," *Frontiers Behav. Neurosci.*, vol. 7, p. 139, Mar. 2013.

[2] F. I. Craik and E. Tulving, "Depth of processing and the retention of words in episodic memory," *J. Experim. Psychol., Gen.*, vol. 104, no. 3, pp. 268–294, 1975.

[3] M. Fan, "How big is the gap and how to narrow it? An investigation into the active and passive vocabulary knowledge of L2 learners," *RELC J.*, vol. 31, no. 2, pp. 105–119, Dec. 2000.

[4] K. Watanabe, T. Sathyanarayana, A. Dengel, and S. Ishimaru, "EnGauge: Engagement gauge of meeting participants estimated by facial expression and deep neural network," *IEEE Access*, vol. 11, pp. 52886–52898, 2023.

[5] K. Watanabe, A. Dengel, and S. Ishimaru, "Metacognition-EnGauge: Real-time augmentation of self-and-group engagement levels understanding by gauge interface in online meetings," in *Proc. Augmented Hum. Int. Conf.*, New York, NY, USA, Apr. 2024, pp. 301–303.

[6] G. S. O'Keeffe and K. Clarke-Pearson, "The impact of social media on children, adolescents, and families," *Pediatrics*, vol. 127, no. 4, pp. 800–804, Apr. 2011.

[7] Y. Hu, L. Manikonda, and S. Kambhampati, "What we Instagram: A first analysis of Instagram photo content and user types," in *Proc. Int. AAAI Conf. Web Social Media*, vol. 8, May 2014, pp. 595–598.

[8] T. B. Brown et al., "Language models are few-shot learners," in *Proc. 34th Int. Conf. Neural Inf. Process. Syst.*, Red Hook, NY, USA, Jan. 2020, pp. 1–14.

[9] F. I. M. Craik and R. S. Lockhart, "Levels of processing: A framework for memory research," *J. Verbal Learn. Verbal Behav.*, vol. 11, no. 6, pp. 671–684, Dec. 1972.

[10] J. D. Bransford and M. K. Johnson, "Contextual prerequisites for understanding: Some investigations of comprehension and recall," *J. Verbal Learn. Verbal Behav.*, vol. 11, no. 6, pp. 717–726, Dec. 1972.

[11] E. Tulving and S. Madigan, "Memory and verbal learning," *Annu. Rev. Psychol.*, vol. 21, no. 1, pp. 437–484, Jan. 1970.

[12] T. Rogers, N. A. Kuiper, and W. S. Kirker, "Self-reference and the encoding of personal information," *J. Pers. Soc. Psychol.*, vol. 35, no. 9, pp. 677–688, Jan. 1977.

[13] C. Jacquet, O. Augereau, N. Journet, and K. Kise, "Vocabulometer, a web platform for ubiquitous language learning," in *Proc. ACM Int. Joint Conf. Int. Symp. Pervasive Ubiquitous Comput. Wearable Comput.*, New York, NY, USA, Oct. 2018, pp. 361–364.

[14] K. Yamaguchi, M. Iwata, A. Vargo, and K. Kise, "Mobile Vocabulometer: A context-based learning mobile application to enhance English vocabulary acquisition," in *Proc. Adjunct ACM Int. Joint Conf. Pervasive Ubiquitous Comput. Proc. ACM Int. Symp. Wearable Comput.*, New York, NY, USA, Sep. 2020, pp. 156–159.

[15] A. Hautasaari, T. Hamada, K. Ishiyama, and S. Fukushima, "VocaBura: A method for supporting second language vocabulary learning while walking," *Proc. ACM Interact., Mobile, Wearable Ubiquitous Technol.*, vol. 3, no. 4, pp. 1–23, Dec. 2019.

[16] R. Arakawa, H. Yakura, and S. Kobayashi, "VocabEncounter: NMT-powered vocabulary learning by presenting computer-generated usages of foreign words into users' daily lives," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, New York, NY, USA, Apr. 2022, pp. 1–21.

[17] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," *ACM Comput. Surv.*, vol. 55, no. 12, pp. 1–38, Mar. 2023.

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., Jun. 2017, pp. 5998–6008.

[19] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling laws for neural language models," 2020, *arXiv:2001.08361*.

[20] A. Chowdhery et al., "PaLM: Scaling language modeling with pathways," *J. Mach. Learn. Res.*, vol. 24, no. 240, p. 113, Jan. 2022.

[21] J. W. Rae et al., "Scaling language models: Methods, analysis & insights from training gopher," 2021, *arXiv:2112.11446*.

[22] Y. Park, G. T. LaFlair, Y. Attali, A. Runge, and S. Goodwin, "Interactive reading-the Duolingo English test," Duolingo Res., Pittsburgh, PA, USA, White Paper DDR-22-02, 2022. [Online]. Available: https://doi.org/10.46999/rcxb1889

[23] J. Leong, P. Pataranutaporn, V. Danry, F. Perteneder, Y. Mao, and P. Maes, "Putting things into context: Generative AI-enabled context personalization for vocabulary learning improves learning motivation," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, New York, NY, USA, May 2024, pp. 1–15.

[24] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3156–3164.

[25] L. Reynolds and K. McDonell, "Prompt programming for large language models: Beyond the few-shot paradigm," in *Proc. Extended Abstr. CHI Conf. Hum. Factors Comput. Syst.*, New York, NY, USA, May 2021, pp. 1–7.

[26] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," 2021, *arXiv:2106.09685*.

[27] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "LLaMA: Open and efficient foundation language models," 2023, *arXiv:2302.13971*.

[28] G. Team et al., "Gemma: Open models based on Gemini research and technology," 2024, *arXiv:2403.08295*.

[29] K. Yamaoka, K. Watanabe, K. Kise, A. Dengel, and S. Ishimaru, "Experience is the best teacher: Personalized vocabulary building within the context of Instagram posts and sentences from GPT-3," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, Sep. 2022, pp. 1–13.

[30] E. Toch, Y. Wang, and L. F. Cranor, "Personalization and privacy: A survey of privacy risks and remedies in personalization-based systems," *User Model. User-Adapted Interact.*, vol. 22, nos. 1–2, pp. 203–220, Apr. 2012.

[31] S. Y. Y. Chyung, M. Kennedy, and I. Campbell, "Evidence-based survey design: The use of ascending or descending order of Likert-type response options," *Perform. Improvement*, vol. 57, no. 9, pp. 9–16, Oct. 2018.

[32] R. Speer. (Sep. 2022). *Rspeer/wordfreq: V3.0*. [Online]. Available: https://doi.org/10.5281/zenodo.7199437

[33] K. Knight and J. Graehl, "Machine transliteration," 1997, *arXiv: cmp-lg/9704003*.

[34] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. 3rd Int. Conf. Learn. Represent.*, San Diego, CA, USA, Y. Bengio and Y. LeCun, Eds., Jan. 2015, pp. 1–5.

[35] J.-B. Alayrac et al., "Flamingo: A visual language model for few-shot learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., Jan. 2022, pp. 23716–23736.

[36] J. Wang, Z. Yang, X. Hu, L. Li, K. Lin, Z. Gan, Z. Liu, C. Liu, and L. Wang, "GIT: A generative image-to-text transformer for vision and language," 2022, *arXiv:2205.14100*.

[37] OpenAI et al., "GPT-4 technical report," 2023, *arXiv:2303.08774*.

[38] L. Beyer et al., "PaliGemma: A versatile 3B VLM for transfer," 2024, *arXiv:2407.07726*.

[39] C. Morikawa, M. Kobayashi, M. Satoh, Y. Kuroda, T. Inomata, H. Matsuo, T. Miura, and M. Hilaga, "Image and video processing on mobile devices: A survey," *Vis. Comput.*, vol. 37, no. 12, pp. 2931–2949, Dec. 2021.

[40] G. Team et al., "Gemini: A family of highly capable multimodal models," 2023, *arXiv:2312.11805*.

KANTA YAMAOKA was born in Hikari, Yamaguchi, Japan, in 2000. He received the B.E. degree from Osaka Prefecture University, Japan, in 2023. From 2022 to 2023, he was a Research Assistant with German Research Center for Artificial Intelligence (DFKI), Germany. Also, he was an Exchange Student with RPTU Kaiserslautern-Landau (formerly known as University of Kaiserslautern), Germany, from 2022 to 2023, which was supported by JASSO study abroad scholarship. In 2023, he joined one of the leading companies in the IT industry, where he contributes to the maintenance and deployments of hyperscale computing infrastructures that are mission-critical to society. Aside from his profession, he has been working on his personal research projects with his personal hours, exploring creative and practical applications of LLMs in the EdTech realm.

KO WATANABE was born in Hiroshima, Japan, in 1994. He received the B.E. degree in mechanical engineering from Tokyo University of Agricultural and Technology, Japan, in 2017, the M.E. degree from Nara Institute of Science and Technology, Japan, in 2019, and the Ph.D. degree in computer science from RPTU Kaiserslautern-Landau, in 2024. He was a Software Engineer with DeNA, Tokyo, Japan. His current research interests include technologies that augment human intellect and fairness in medical AI.

KOICHI KISE (Member, IEEE) received the B.E., M.E., and Ph.D. degrees in communication engineering from Osaka University, Osaka, Japan, in 1986, 1988, and 1991, respectively. From 2000 to 2001, he was a Visiting Professor with German Research Center for Artificial Intelligence (DFKI), Germany. He is currently a Professor with the Department of Computer Science and Intelligent Systems and the Director of the Institute of Document Analysis and Knowledge Science (IDAKS), Osaka Prefecture University, Japan. His major research interests include analysis, recognition and retrieval of documents, images, and activities. He is also a member of IEICE, ACM, IPSJ, IEEJ, ANLP, and HIS. He was a recipient of the Best Paper Award of IEICE, in 2006, the IAPR/ICDAR Best Paper Awards, in 2007 and 2013, the IAPR Nakano Award, in 2010, the ICFHR Best Paper Award, in 2010, and the ACPR Best Paper Award, in 2011. He worked as the Chair of the IAPR Technical Committee 11 (reading systems) and a member of the IAPR Conferences and Meetings Committee. He is currently the Editor-in-Chief of the *International Journal of Document Analysis and Recognition*.

ANDREAS DENGEL received the Diploma degree in CS from TUK and the Ph.D. degree from the University of Stuttgart. He is currently a Scientific Director of DFKI GmbH, Kaiserslautern. In 1993, he became a Professor in computer science with TUK, where he holds the Chair Knowledge-Based Systems. Since 2009, he has been appointed as a Professor (Kyakuin) with the Department of Computer Science and Information Systems, Osaka Prefecture University. He was also with IBM, Siemens and Xerox Parc. He is a member of several international advisory boards, has chaired major international conferences, and founded several successful start-up companies. He is a co-editor of international computer science journals and has written or edited 12 books. He is the author of more than 300 peer-reviewed scientific publications and supervised more than 170 master's and Ph.D. theses. He is a fellow of IAPR and received many prominent international awards. His main scientific emphasis is in the areas of pattern recognition, document understanding, information retrieval, multimedia mining, semantic technologies, and social media.

SHOYA ISHIMARU (Member, IEEE) was born in Ehime, Japan, in 1991. He received the B.E. and M.E. degrees in electrical engineering and information science from Osaka Prefecture University, Japan, in 2014 and 2016, respectively, and the Ph.D. degree (summa cum laude) in engineering from the University of Kaiserslautern, Germany, in 2019. He has been a Project Professor with the Department of Computer Science, Osaka Metropolitan University, Japan, since 2023. In addition, he has been an Associate Director of Japan Laboratory of German Research Center for Artificial Intelligence (DFKI Laboratory), Japan, since 2023, and a Researcher with the Keio Media Design Research Institute, since 2014. He was a Junior Professor with the University of Kaiserslautern-Landau, Germany, from 2021 to 2023, and was a Senior Researcher with DFKI, from 2019 to 2023. His research interests include human–computer interaction, machine learning, and cognitive psychology with the aim of amplifying human intelligence. His awards and honors include the Best Presentation Award at Asian CHI Symposium, in 2020, Poster Track Honorable Mention at UbiComp/ISWC, in 2018, and MITOU Super Creator, which is a title given to outstanding software developers (around ten people per year) by the Ministry of Economy, Trade, and Industry in Japan.