Putting Things into Context: Generative AI-Enabled Context Personalization for Vocabulary Learning Improves Learning Motivation

Joanne Leong
MIT Media Lab
Cambridge, MA, USA
joaleong@media.mit.edu

Pat Pataranutaporn
MIT Media Lab
Cambridge, MA, USA
patpat@mit.edu

Valdemar Danry
MIT Media Lab
Cambridge, MA, USA
vdanry@mit.edu

Florian Perteneder
Independent
Hagenberg, Austria
floperteneder@gmail.com

Yaoli Mao
Columbia University
New York, NY, USA
ym2429@tc.columbia.edu

Pattie Maes
MIT Media Lab
Cambridge, MA, USA
pattie@media.mit.edu

ABSTRACT

Fostering students' interests in learning is considered to have many positive downstream effects. Large language models have opened up new horizons for generating content tuned to one's interests, yet it is unclear in what ways and to what extent this customization could have positive effects on learning. To explore this novel dimension, we conducted a between-subjects online study (n=272) featuring different variations of a generative AI vocabulary learning app that enables users to personalize their learning examples. Participants were randomly assigned to control (sentence sourced from pre-existing text) or experimental conditions (generated sentence or short story based on users' text input). While we did not observe a difference in learning performance between the conditions, the analysis revealed that generative AI-driven context personalization positively affected learning motivation. We discuss how these results relate to previous findings and underscore their significance for the emerging field of using generative AI for personalized learning.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; • **Computing methodologies** → *Natural language generation*; • **Applied computing** → *Education*.

KEYWORDS

generative artificial intelligence, education, learning, vocabulary

ACM Reference Format:
Joanne Leong, Pat Pataranutaporn, Valdemar Danry, Florian Perteneder, Yaoli Mao, and Pattie Maes. 2024. Putting Things into Context: Generative AI-Enabled Context Personalization for Vocabulary Learning Improves Learning Motivation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3613904.3642393

CHI '24, May 11–16, 2024, Honolulu, HI, USA
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642393

1 INTRODUCTION

Catering learning experiences to students' interests is considered to have many positive effects. It can make students more persistent [4], and it impacts students' career choices [19]. Yet, according to Hidi and Harackiewicz [20], motivating students is one of the greatest challenges in education. One impediment to fostering interest is the one-size-fits-all approach to educational materials. While this has had a crucial benefit of making education accessible to a large number of people, it can often fall short in engaging students. Since each student arrives with different prior knowledge and interests, some students have trouble connecting with these materials and consequently lose interest in learning [30].

In light of this, the arrival of landmark AI developments is considered an opportunity to revolutionize education, bringing new hopes for exciting educational experiences. In particular, the introduction of advanced large language models (LLMs) such as GPT-3/4 and ChatGPT,[1] which are capable of dynamically generating passages of text that are difficult to distinguish from authentic human writing, opens up numerous possibilities. On one hand, generative AI in education raises many concerns around cheating and rising unemployment in the education sector [46]. On the other hand, it unlocks new possibilities to support learning [32]. For instance, Khan Academy[2] and Duolingo[3] are exploring how generative AI could drive tailored tutoring systems. Other works are exploring how AI could facilitate learning via reading and writing stories [45], or creating custom picture flashcards [12].

We note that another exciting possibility is to leverage generative AI for context personalization [47], wherein learning content and/or tasks are customized to the individual student. Approaches to this commonly aim to incorporate learners' individual preferences and interests into materials, and previous educational studies have provided some evidence that doing this can boost students' situational interest, value perception, and task effort when learning [22, 23]. Despite this, investigations in this space have been relatively few. This is understandable, since manually creating materials that can be adapted to accommodate the varied interests of a large number

[1] https://openai.com/research/overview
[2] https://www.khanacademy.org/khan-labs
[3] https://blog.duolingo.com/duolingo-max/

of students is an immensely effortful and time-consuming undertaking. In light of this, generative AI can dynamically produce written text that adapts to given input, making it a promising candidate technology for enabling personalization at scale. Yet, to our knowledge, using text-to-text based generative AI to tackle this challenge has not yet been formally investigated, and the practical implications for using it for this use case are unknown.

Therefore, it is at this intersection of education and technology that we situate our research. Amidst both great enthusiasm and concern in this space, we take a pause to thoroughly examine generative AI's potential to induce improvements in learning. Specifically, our goal in this work is to gain insight into the approach people may take to leveraging this technology to create individualized learning examples, as well as to investigate its merits for improving learning outcomes and enhancing the overall learning experience. In light of numerous possibilities, we focus on advanced vocabulary learning for native (L1) and fluent non-native English language learners (L2). To do this, we developed a web-based vocabulary learning app as a technology probe that supports different modes of learning, and conducted a large between-subjects experiment with 272 participants online. Participants were randomly assigned to one of three conditions. In each condition, participants were given a series of new vocabulary words and their definitions, alongside a particular example for how the word could be used in text, dependent on their assigned condition. In the **control** condition, participants received a sentence quoted from a pre-existing book or article. In the **generated-sentence** condition, participants typed in an input to receive a personally contextualized and dynamically generated sentence, and in the **generated-story** condition, participants typed in an input to receive a personally contextualized dynamically generated short story. Following this learning experience, participants were asked to answer a set of survey questions and complete a vocabulary quiz. One week later, participants were asked to retake the same quiz. Analyzing the results, we found no indication of a difference in learning performance between the control and the generative AI-enabled personally contextualized learning conditions. However, we found strong evidence that the context personalization of learning materials can enhance people's perception of the learning experience, including boosting their levels of intrinsic motivation. We also discovered that people used different strategies when providing inputs to steer the context personalization.

In summary, the contributions of this work include (1) an outline of different strategies people may use to personalize text-based learning materials and their underlying motivations; (2) evidence that generative AI-enabled context personalization can enhance people's appraisal of their learning experience and feelings of motivation, as derived from a large online between-subjects study (n=272); and (3) a set of implications derived from the study findings that can inform future research and development towards facilitating personalized learning experiences with generative AI.

2 BACKGROUND & RELATED WORK

There is a rich body of work pertaining to personalization in education. To underpin our investigation on utilizing generative AI for context personalization, we first present perspectives and theories that consider personalization as an asset to foster learning motivation. Second, we highlight the dimensions of context personalization and the different levels of contextualization, and finally discuss prior context-based vocabulary learning tools.

2.1 Motivating Students via Personalization

Over the last decade, the interest in personalized education has been on the rise. In 2016, the U.S. Department of Education Office of Educational Technology brought a focus to personalized learning (PL), defining it as "*...instruction in which the pace of learning and the instructional approach are optimized for the needs of each learner. Learning objectives, instructional approaches, and instructional content (and its sequencing) all may vary based on learner needs. In addition, learning activities are meaningful and relevant to learners, driven by their interests, and often self-initiated.*" [54, p. 9]

In other words, personalized learning is learner-centric, where instructional content is focused on addressing each learner's needs and interests [5]. There are multiple approaches to personalizing learning. One type of personalization that has been implemented is **context personalization** [47]. It involves drawing upon and grounding information in the context of individuals' interests (such as sports, music, or video games), their preferences (such as favorite foods), personal information (such as the name of their friends and pets), or their "existing funds of knowledge" [58]. Several studies show that context personalization can help to motivate students and boost their level of achievement [22, 23, 31, 36, 44, 58, 59]. Context personalization has previously been implemented by surveying students and injecting keywords [22, 23] and personal information [36] into their learning materials. This principle has also been applied outside of textual learning materials, for instance by substituting video instructors with generated-characters resembling figures that students like and admire [44].

All these efforts on personalization are driven by the goal to foster students' interest, which was identified to be an important condition for learning that has downstream effects on students' attention, goals, and levels of learning [21, 48]. Interest can either refer to the psychological state of engagement, or the predisposition of a person to reengage with a topic or idea over time. Although interest could be construed as something that students either have or do not have, Hidi and Renninger [21] have proposed that interest is something that can be shaped and developed over time. They proposed a four-phased model of interest, with the first two phases encapsulating the triggering and maintenance of **situational interest** and the last two phases covering the emergence of and maturation of **individual interest**. According to their model, it is necessary to first generate situational interest before students can develop a more stable individual interest in a topic. Most interventions traditionally tackle the first two stages of interest. Since students come from diverse backgrounds, Reber et al. [47] posited that personalized materials have a higher probability of increasing situational interest than a standardized, one-size-fits-all approach. In other words, context personalization may be a means to heighten the psychological state of situational interest.

Our aim to examine the utility of generative AI for context personalization marks a departure from prior literature on the topic, which has to date relied on manually crafted examples [59] or preprepared 'fill-in-the-gaps' templates in which students' details are injected [22, 23, 31, 36]. These approaches have necessitated surveys or interviews to be conducted prior to the main learning activity, and have yielded rigid materials, which for the most part only superficially mention individual students' details. This requires extensive effort from educators to anticipate and draw relationships between target topics and students' details, making it challenging to scale. Generative AI in contrast is well-positioned to overcome these hurdles as it can interactively accept inputs directly from learners and dynamically produce text about target concepts with respect to virtually any topic – almost instantly and at scale. Paired with its user-friendliness, these attributes position generative AI advantageously for broad future adoption and make it well-suited for moving beyond a one-size-fits-all model of teaching. However, as its application for this use case has not previously been explored, its potential and limitations in the context of personalizing learning materials needs to be investigated for developing best practices. With this work, we aim to contribute to this understanding and pave the way for deeper investigations into facilitating personalized learning experiences with generative AI.

2.2 Dimensions of Context Personalization

Walkington and Bernacki [57] identified multiple dimensions for context personalization: *depth of personalization, grain size, and ownership*. **Depth of personalization** refers to what degree the materials integrate with the learner's interest. Shallow connections introduce a student's name or preferences into the learning material in a fill-in-the-blank manner. In contrast, deep connections infuse the larger context of the given topic, for instance by crafting an entire sentence around the given preferences. A study by Høgheim and Reber [23] compared these two levels and found that the latter was superior. **Grain size** refers to whether the material caters to the overall interests of a group of people, or the interests of a particular individual. In a study by [36], it was found that more individual personalization prompted more interest. Another factor is **ownership**. Personalization can be carried out by teachers or curriculum designers or students themselves can play an active role in personalizing their learning materials. Ownership is also closely connected to **autonomy**, which was investigated by Frenzel et al. [16]. Their work highlighted that especially as children transition into adolescence, they have a growing need for autonomy. They recommend supporting this cognitive need by offering choices in learning activities, for instance, by providing options to shape the nature of the instructions or questions. In prior studies, providing students with the opportunity to choose and personalize their questions based on popular figures, places, and themes [22, 23] was found to be particularly helpful for lower scoring students in mathematics. We see generative AI as well-positioned to perform deep personalization, support individual grain size at scale, and enable student ownership of the generation of learning materials. In our work we pay attention to these dimensions and aim to address them in our design choices.

2.3 Levels of Contextualization for Learning

In the language learning domain, contextualization can be interpreted not only in a thematic sense, but also in a functional sense. For example, the location of a word within a passage (a spatial cue) and how it relates to other words (a functional cue) can help users to guess the meaning of a foreign word. As such, Oxford and Scarcella [41], studying second language vocabulary instruction, identified three types of activities: *decontextualized, partially contextualized, and fully contextualized*. Decontextualized activities are those that isolate the vocabulary word from meaningful context (e.g., word lists, flashcards). Partially contextualized activities give slightly more context (e.g., by presenting word groups, adding visual/auditory information, etc.) Lastly, fully contextualized activities present words within engaging, meaningful, and authentic communication scenarios via reading, listening, speaking etc. With our experimental conditions, we aim to approach higher levels of contextualization, as recommended by the authors.

Fully contextualized learning can be achieved in many ways, including reading stories. Some research in neuroscience corroborates this idea. For instance, some studies have found that incorporating social elements can significantly aid memory encoding compared to non-social memorization [11, 18, 28], and may trigger increased learning effects, even in subjects like math and science [35]. Other research underlines the motivational benefits that storytelling and narratives have on learning [39, 40]. On this basis, numerous educational initiatives are turning towards developing interactive story-based learning materials [8, 62]. While these approaches take advantage of the benefits of narrative learning, they often require significant resources and time to develop. With the advent of new AI developments, generative models can alleviate this challenge by creating personalized and dynamic narratives for each student.

2.4 Context-Based Vocabulary Learning Tools

Many prior systems have been designed to help support people in learning new vocabulary words and grammar in both real-world and virtual contexts. For example, several computer vision and AR-based tools have been designed to enable people to learn vocabulary and grammar in-situ, such as by overlaying labels on real-world objects [13, 24, 56, 61]. Other systems have integrated vocabulary learning with people's media consumption. Smart Subtitles [29] was an interactive system that enabled people to learn words from a foreign language while watching videos, whereas Vázquez et al. [55] created a VR app that paired the user's performance of an action with the pronunciation of that verb in a foreign language. Lungu et al. presented a language textbook prototype that enabled foreign language learners to read materials that are personally interesting to them from the web [37], which received positive feedback from the students and teachers. In more recent years, new techniques to contextualize learning have emerged with advancements in natural language processing (NLP) and generative AI. VocabEncounter by Arakawa et al. [7], was a browser extension that could embed foreign vocabulary words into a user's web content, allowing them to review target words in the context of web browsing. Draxler et. al [12] created a mobile app for children, which generated multi-media learning materials based on photos taken on their phone. In a similar vein, Draxler et al. conducted a study comparing personalized auto-generated flashcards against crowdsourced ones, and found that learning decreased with personalization [14]. Yamaoka et al. [63] prototyped a system that used GPT-3 to generate sentences based on a learner's image-based posts to Instagram in order to expose them to new vocabulary in a foreign language related to their personal experiences. Storyfier [45] was a prototype system to aid people in learning a set of vocabulary words through reading generated stories containing the words, cloze (i.e., fill-in-the-blank) tests with the content, and co-writing with AI. In addition to their approach of embedding vocabulary in a narrative, we are interested in understanding what happens when that is coupled with thematic contextualization framed around people's interests. Most of the other works discussed in this section try to enable people to learn words in their physical or online context. Our work tries to study the implications of allowing people to specify the context in which the information is consumed, which can be unrelated to their current situated context but rather more related to their existing interests and funds of knowledge.

3 STUDY CONCEPT & RATIONALE

The goal of this work is to investigate the approach people take towards using generative AI to personally contextualize their own learning examples, and to explore its merits for improving learning outcomes and enhancing the overall learning experience. Our exploration centers on three main areas of interest. Firstly, how might learners approach user-driven context personalization opportunities, and how do they perceive AI-generated examples? Secondly, how does context personalization influence learning outcomes? Finally, how does context personalization in these forms shape attitudes and emotions towards learning?

To this end, we aim to develop an app as a technology probe to explore these questions, rather than to design a finalized app for end users. Recalling the design dimensions outlined by Walkington et al. [58], we target an app that enables *deep levels of personalization* (by not only injecting user interests into learning examples e.g., via keyword insertion, which only changes the context on the surface level, but by generating content that more wholly plays on this information), by affording individualized personalization, and by supporting learner agency in the learning process.

To approach this research, we firstly scope our exploration space. While AI can be used to generate many forms of media including photos and videos, we restrict our investigation to **text-based mediums**. Additionally, while there are a large number of topics that can be taught, we opt for a **vocabulary learning task**. This task was chosen since it does not require prior specific domain knowledge, and it is a scenario for which we can relatively simply and feasibly attempt to quantify and measure the degree of learning.

In order to explore the merits of generative AI for context personalization, we designed **three conditions** for the app to compare with one another. As a baseline condition, we took what is the standard experience of learning a new word in one's native tongue (where translation is not an option) online; a vocabulary word and its definition would be presented, alongside an example of how the word can be used in a sentence. The sentence appears without any option for the user to adapt it. As an experimental condition, we decided to enable users to personalize the learning example, on-the-fly, based on a word or phrase provided by them. In order to elicit input from the user, we provided a text box with the instruction "*Generate an example usage of the word [vocabulary word] based on*:" and in the field, they were encouraged to input "*a topic or theme of interest*." In addition, given prior research that suggests that stories are powerful mediums through which to learn [39, 40], we added one more experimental condition that would allow people to generate stories, rather than only a single sentence, tailored to their inputs. In summary, the conditions in our study (pictured in Figure 1) were as follows:

(1) **Control Condition:** participants see example sentences sourced from pre-existing articles and books.
(2) **Generated-Sentence (Gen-Sentence) Condition:** participants see personally contextualized AI-generated sentences based on prompts that incorporate their typed inputs.
(3) **Generated-Story (Gen-Story) Condition:** participants see personally contextualized AI-generated short stories based on prompts that incorporate their typed inputs.

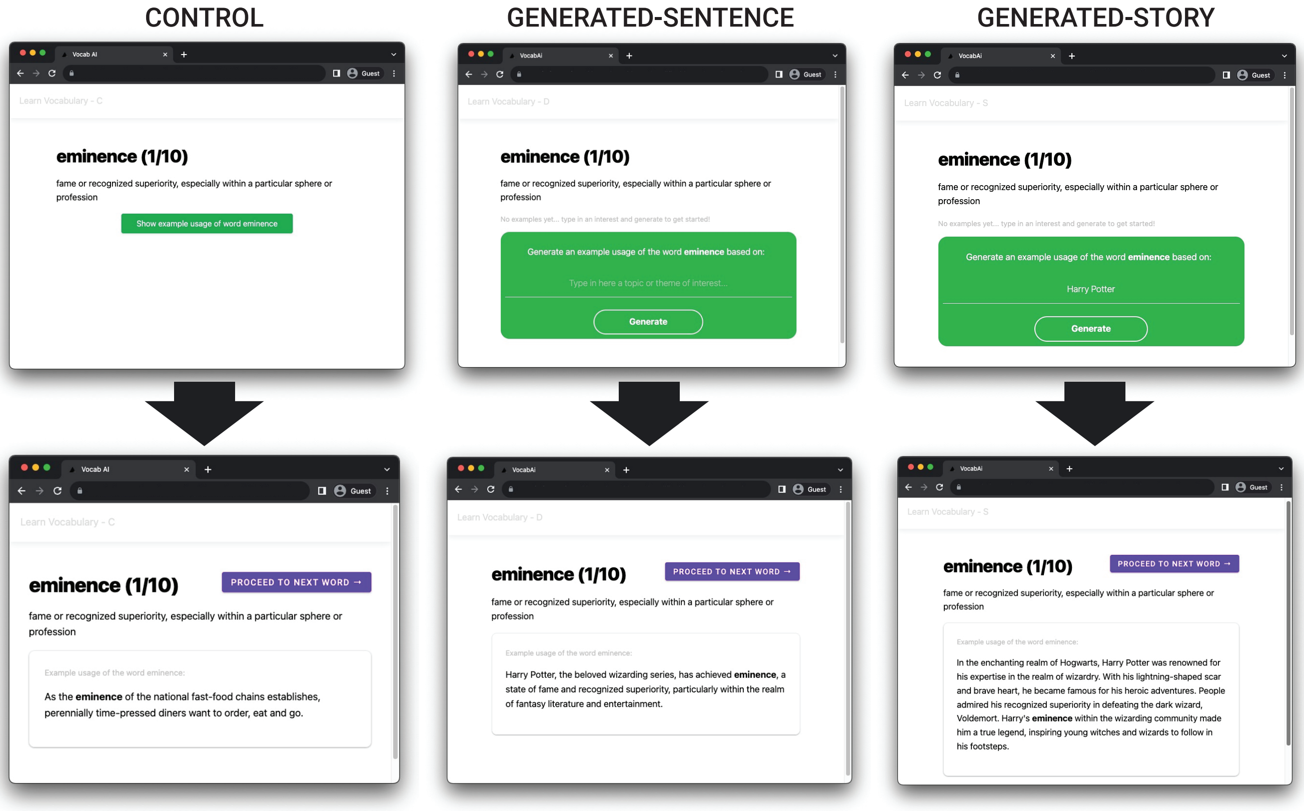
We followed a strict study and app design such that these conditions could be compared fairly to one another on the basis of the factors of interest. Whereas Gen-Sentence is used to probe the impact of context personalization, Gen-Story is used to explore the impact of context personalization coupled with narrative. To isolate for these effects, several limitations were then applied to the app design. Firstly, users were limited to seeing one example usage per word across all conditions. Additionally, participants could only move forwards through the lesson and could not backtrack to review words. Since rehearsal and particularly spaced recall is known to be a helpful learning technique [17], we anticipated users would differ from one another along this behavior. Hence, we eliminated this option to remove its influence. Finally, examples were chosen and generated to encompass only one instance of the word, rather than multiple. These limits helped to ensure that any effects could be attributed to the conditions, rather than the number of exposures participants had to vocabulary word information.

With regards to eliciting user input, we had the option to either ask users to answer a survey at the beginning of the overall experience to gauge their interests, or to provide an input prior to each example. We decided on the latter design option. Using this approach not only gives users more autonomy over how their learning experience is personalized (by implicitly choosing which input maps with which vocabulary word), but it also allows patterns of behavior to organically emerge.

4 SYSTEM DESIGN & IMPLEMENTATION

In order to facilitate our research exploration, we designed and implemented a vocabulary web app that enables three different modes of learning. Taking a set of vocabulary words that the user has declared as not knowing, it presents information about the words in sequence with each word being presented on its own page with its definition. Additionally, depending on the condition to which the user was randomly assigned, either a button or a text box would be available to show or generate an example of how the word could be used in text, respectively. In the control condition, a user presses a button to reveal an example that is sourced from existing articles and books, as found on the Cambridge Dictionary website.[4] In the gen-sentence condition, the user is instructed in the interface to type in a topic or theme of interest. Their input is then integrated into a prompt and passed to a generative AI model

[4] https://dictionary.cambridge.org/us/dictionary/english/



eminence (1/10)

fame or recognized superiority, especially within a particular sphere or profession

No examples yet... type in an interest and generate to get started!

Generate an example usage of the word **eminence** based on:

Type in here a topic or theme of interest...

Harry Potter

Example usage of the word eminence:

As the **eminence** of the national fast-food chains establishes, perennially time-pressed diners want to order, eat and go.

Harry Potter, the beloved wizarding series, has achieved **eminence**, a state of fame and recognized superiority, particularly within the realm of fantasy literature and entertainment.

In the enchanting realm of Hogwarts, Harry Potter was renowned for his expertise in the realm of wizardry. With his lightning-shaped scar and brave heart, he became famous for his heroic adventures. People admired his recognized superiority in defeating the dark wizard, Voldemort. Harry's **eminence** within the wizarding community made him a true legend, inspiring young witches and wizards to follow in his footsteps.

Figure 1: Examples for the three variations of the vocabulary app as they would appear for the vocabulary word “eminence.” In the control condition, participants would click a button to show an example usage of the target word, sourced from a pre-existing article. In the generated-sentence and generated-story conditions, users would enter an input (in this depiction, “Harry Potter,”) to generate a custom learning example. After the example is shown, a button would appear to allow participants to proceed to learn the next vocabulary word. Note that the first screenshot for the generated-sentence condition depicts how the textbox would look prior to a typed entry.

(OpenAI’s GPT-turbo-3.5) to dynamically create an example. The user experience for the gen-story condition was identical to the gen-sentence condition, except that it would output a story for the user corresponding to their typed input, based on a different underlying prompt. The user interface (UI) for each condition, as pictured in Figure 1, was designed to be as similar to one another as possible to make the conditions more cleanly comparable based on the learning example given, rather than the interface. After the corresponding example was revealed, a button would appear in the top-right hand corner to allow the user to proceed to the next vocabulary word page.

The web application was implemented using NextJS[5] and deployed via Vercel.[6] The app calls OpenAI’s API in order to generate text. Upon revealing an example, the results were logged into a Google Sheet via its corresponding API.

An iterative approach was used to design the AI prompts. Special efforts were made to ensure the quality of the generated examples. For instance, the generated text should contain the vocabulary word and be within a certain word length. Furthermore, we wanted the generated sentences and stories to not only contain the target word but to also incorporate its definition to reinforce its meaning. This was done to steer the generation towards showcasing the meaning of the vocabulary word as would traditional examples meant to highlight its use.

The prompt for the generated-sentence condition was as follows:

> *Write a medium or long-length sentence that contains the word “[vocabulary-word]” with its meaning “[definition]” related to the topic of “[user-input]”. It must contain exactly the word “[vocabulary-word]”, and it must also contain a paraphrasing of the meaning of the word “[vocabulary-word]” within a nonessential clause.*

In contrast, a chained prompt was used to generate stories in the gen-story condition. The first prompt was as follows:

> *Definition of [vocabulary-word]:[definition]. Write a short children’s story framed around the word “[vocabulary-word]”. It has to be about [user-input]. Make sure the word “[vocabulary-word]” is used and defined within a*

[5] https://nextjs.org/
[6] https://vercel.com/

Table 1: Real learning examples that were displayed for the vocabulary word "predilection" under each study condition. The AI-generated learning examples were created based on the user input "animals." Font styling is used here for greater legibility.

Condition	Vocab. Word	User Input	Learning Example
Control	***predilection***	n/a	Ever since she was a child, she has had a ***predilection*** for spicy food.
Gen-Sentence	***predilection***	animals	Despite her ***predilection*** for cats, which can be defined as a strong preference or special liking for felines, Sarah's unbiased love for all animals led her to volunteer at the local animal shelter, where she cared for dogs, rabbits, and even reptiles with equal devotion.
Gen-Story	***predilection***	animals	In a lush forest, there lived a wise old owl named Oliver. He had a strong liking for storytelling. Every evening, animals gathered around him, eager to hear his tales. One day, a curious squirrel asked, "Why do you have such a ***predilection*** for storytelling, Oliver?" With a twinkle in his eyes, Oliver replied, "Because stories have the power to ignite our imagination and bring us together."

> *nonessential clause when it is used within the story. (60 words). Do not begin the story with "Once upon a time."*

We noticed in pilot testing that the stories generated by this prompt often contained multiple sentences containing the vocabulary word. Since we only wanted the vocabulary word to appear once in the text (to make this condition comparable to the other conditions), we scanned the output from the first prompt for repetitions. For all sentences that repeated the vocabulary word, we applied a subsequent call to GPT-3.5 turbo with the following prompt:

> *Paraphrase and remove the word '[vocabulary-word]' from this sentence: '[sentence].'*

We then injected this output into the original one to replace the old sentence. While we tried a version of the prompt without specifying a "children's" story, we noticed that the generated outputs would often lack a narrative arc with a clear beginning, middle, and end. As such, we therefore opted to keep this specification in the final prompt design. However, this led to a very high chance of stories starting with "Once upon a time." Since this felt quite repetitive, we chose to include an explicit instruction to avoid this in order to yield more variance in the outputs. Examples of the resulting outputs based on these prompt designs are shown in Table 1.

5 PROCEDURE

To explore the concept of AI-generated learning materials, we conducted an online, two-part, between-subjects study. We embedded our app into Qualtrics [1] and recruited participants using Prolific [3]. This research was approved by the local university's institutional review board (IRB).

Each participant first completed a prescreening and received a $0.15 USD gratuity. This was to ensure that there were at least ten words in our word set that they did not already know and could therefore attempt to learn as part of our study. Eligible participants from the prescreening were then invited to a two-part study procedure. Part 1 took approximately 20 minutes to complete, for which participants were paid $3 USD. Part 2 took approximately 5 minutes to complete, for which participants were compensated an additional $2 USD. The details for each phase are visualized in Figure 2.

5.1 Prescreening

Participants were first asked to complete a prescreening questionnaire on the Prolific platform. Fluency in English was specified as a prescreening criterion within Prolific for access. The pre-screen contained a list of 30 advanced vocabulary words taken from the Graduate Record Examinations (GRE), a standardized test that is a common admissions requirement for graduate schools in the US and Canada. These words were selected from various freely-available online GRE vocabulary word lists, such as from Kaplan,[7] Vocabulary.com,[8] and Magoosh.[9] We asked them to assign a ranking to each word based on a 4-point adapted version of the Vocabulary Knowledge Scale (VKS) [10] (1 = "*I have never seen this word before, and do not know what it means*", 4 = "*I know what this word means, and I can use it in a sentence*"), based on the original VKS [43]. A list of 10 words was created for each participant based on words they rated as 1 on the VKS. If the participant did not mark a total of 10 or more words that met this condition, the participant was not eligible to participate in the full study.

5.2 Part 1 - Lesson, Survey & Vocabulary Quiz

Eligible participants from the prescreening were invited to Part 1 of the study. After reading an overview of the study and providing their consent, participants were randomly assigned to one of the three conditions and were directed to the respective version of the vocabulary learning web app. The app presented each participant with a set of vocabulary comprising the first 10 words that they had assigned a VKS score of 1 (i.e., the word was completely unfamiliar to them) in the prescreening phase. Following the lesson, participants were asked to answer a set of survey questions related to their perception of the overall experience and to complete a vocabulary quiz. They were then asked to refrain from deliberately studying the vocabulary words they had learned until Part 2. Details of the survey are explained in section 6.

[7] https://www.kaptest.com/study/gre/top-52-gre-vocabulary-words/
[8] https://www.vocabulary.com/lists/18294
[9] https://gre.magoosh.com/flashcards/vocabulary/decks

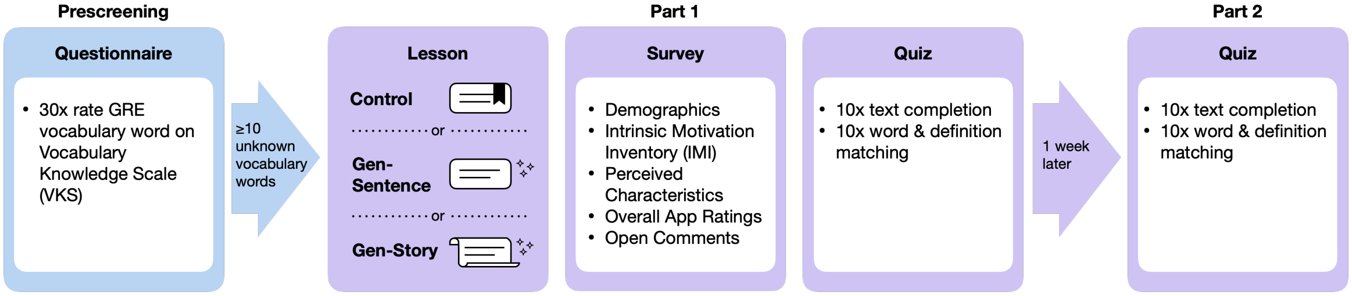

Prescreening
Questionnaire
30x rate GRE vocabulary word on Vocabulary Knowledge Scale (VKS)
≥10 unknown vocabulary words
Part 1
Lesson
Control
or
Gen-Sentence
or
Gen-Story
Survey
Demographics
Intrinsic Motivation Inventory (IMI)
Perceived Characteristics
Overall App Ratings
Open Comments
Quiz
10x text completion
10x word & definition matching
1 week later
Part 2
Quiz
10x text completion
10x word & definition matching

Figure 2: An overview of the user study procedure.

5.3 Part 2 - Repeated Vocabulary Quiz

After one week, participants were invited to complete the same vocabulary quiz. This procedure was modeled after previous studies evaluating long-term learning retention [6, 49].

6 MEASURES

The main independent variable in this experiment was the condition i.e., ***control***, *generated-sentence (**gen-sentence**)*, and *generated-story (**gen-story**)*. The dependent variables were selected in alignment with common practices in prior studies on context personalization [47], and were extended to encompass measures that could help probe how AI-generated outputs may be perceived in this use case.

App Logs: The vocabulary app logged participants' typed inputs in the gen-sentence and gen-story conditions, the provided example, and the timestamp for when the participant clicked the button to show or generate the example.

Learning Performance: This grouping pertains to participants' acquisition and retention of learning content from their learning conditions. Each participant completed a vocabulary quiz at two distinct time-points, once after the lesson, and again 1 week later. The quiz comprised a total of 20 multiple-choice questions. Ten questions were based on existing **GRE Verbal Text Completion and Sentence Equivalence questions**. These questions were pulled from existing GRE practice exams from various sources, including the Princeton Review,[10] MainTests,[11] Kaplan,[12] and Varsity Tutors.[13] The other ten questions required participants to choose the appropriate definition to match a word, or vice versa. All the options that were provided were based on an original set of 30 GRE words, and were matched to be of the same type (i.e., noun, verb or adjective). Participants chose the correct answer from five available options. Questions were presented to participants in a randomized order. We decided to refrain from free-recall tests, since it is possible that given a definition, a participant could provide a synonym to the intended word.

Learning Experience: This grouping captures participants' subjective experience of their learning conditions. Participants were asked to complete a set of survey questions based on the multidimensional **Post-Experimental Intrinsic Motivation Inventory (IMI)** [2, 50]. The full inventory comprises a total of 45 7-point Likert items (1 = *not at all true*, 7 = *very true*), some reverse-coded. The items span seven distinct subscales: *Interest/Enjoyment, Perceived Competence, Effort/Importance, Pressure/Tension, Perceived Choice, Value/Usefulness and Relatedness.* This inventory aims to evaluate participants' subjective experience during a specific laboratory activity in previous research. Despite its name, only the Interest/Enjoyment subscale is utilized as a self-report measure for intrinsic motivation, while the remaining subscales evaluate associated factors. Only factors that are relevant to one's investigation need to be included. Therefore, we removed questions pertaining to *Pressure/Tension and Relatedness.* The wording of some items were adapted to suit the context of our study. Examples of statements included: "*Learning the vocabulary words with this app was fun to do*" (Interest/Enjoyment), and "*Learning new vocabulary words with this app was something I couldn't do very well*" (Perceived Competence, reverse-coded). The Interest/Enjoyment subscale is considered to be the self-report measure of intrinsic motivation. Items included in our survey were presented to participants in a randomized order within each subscale.

Perceived Characteristics of Examples: Participants were asked to rate their perception of the examples they were given in their condition using 7-point Likert items (1 = *not at all*, 7 = *very true*). These were a custom and exploratory set of questions to probe characteristics spanning: *Funny/Humorous, Personally Relatable, Aligned with my Interests, Related to my Knowledge / Subject Matter Expertise, Surprising/Outrageous, Original/Creative, Emotionally Positive, Emotionally Negative.* In addition to probing whether people perceive the examples to be personalized, other aspects were included based on prior works that have suggested that humor and informational incongruity can play a role in learning [60], and that positive emotions may help catalyze academic engagement [15, 27].

Overall App Ratings: Participants were asked to rate how much they would agree (1 = *Strongly Disagree*, 7 = *Strongly Agree*) that "*I would like to use the app again to learn new vocabulary words,*" and "*I would recommend this app to a friend to learn new vocabulary*

[10] https://www.princetonreview.com/grad-school-advice/gre-verbal-practice
[11] https://www.maintests.com/gre/text-completion
[12] https://www.kaplanquizzes.com/review/gre/?success=yes
[13] https://www.varsitytutors.com/gre_verbal-help/text-completion

Table 2: The mean and standard deviations (in parentheses) for various metrics pertaining to the learning examples under each of the learning conditions.

Attribute	Control	Gen-Sentence	Gen-Story
Participant Count	92	89	91
Word Count per Example	16.85 (5.71)	32.05 (6.29)	65.61 (12.02)
Engagement Time (seconds/example)	17.28 (10.26)	38.85 (24.71)	43.97 (23.42)
Flesch-Kincaid Reading Ease	55.99 (22.48)	26.59 (17.63)	71.82 (10.90)

words." These can provide some indication of how much they enjoyed using their given version of the app. They were also asked to report their attitudes towards using the app (1 = *Very Playfully*, 7 = *Very Seriously*).

Open Comments: As part of the survey immediately after the learning experience, participants were asked to comment on their *approach to their user inputs*, *thoughts regarding their learning experience*, *thoughts to improve the app*, and anything else that came to mind. This was to gain richer insight on their perspectives of using the application.

Text Reading Ease: The Flesch-Kincaid Reading Ease metric [26] is widely used to describe how easy it is to comprehend a piece of written text. Its scores range from 0 to 100, with the lowest score representing professional level texts best understood by university graduates, and 100 representing texts that can be understood by an average 5th grade student in the United States.

7 ANALYSIS & RESULTS

In this section, we outline the results of the user study. We explain the demographics of the participants, how they approached contextualizing their learning materials, how they perceived their learning examples, and how it impacted their learning performance and overall learning experience. Our study's main questions, experimental conditions, methodology, power analysis, dependent variables, and data analyses were all pre-registered[14] on AsPredicted. In this investigation, the primary dependent variables (DVs) were learning performance and learning experience measures. These variables were developed based on existing literature, and statistical tests were pre-registered for their analysis. Additionally, we explored secondary variables. We examined how participants perceived their learning examples across different conditions. We also provided descriptive statistics and qualitative data to depict how participants contextualized their learning materials and their overall user experience with the app. A total of 712 people completed the prescreening. 442 of them met the requirements and were invited to the study. Of those, 301 completed Part 1 (1 response was removed for failing both attention checks, 4 removed for missing IDs), and 288 completed Part 2 (16 responses were excluded for passing the 3 day expiration date of their invitation), resulting in 272 complete responses. A prior simulation-based power-analysis based on 12 samples per condition indicated that a minimum of 180 participants would be needed to reach 80% power. Therefore our study was sufficiently powered.

7.1 Participants

Out of the 272 participants, 67 were between 18-24 years old, 122 were between 25-34, 53 were between 35-44, 15 were between 45-54, 10 were between 55-64, 3 were above 65 years of age, and 2 did not disclose. 144 identified as women, 125 identified as men, 3 identified as non-binary / third gender, and 1 person preferred not to say. Their education levels varied. 69 completed high school, 130 completed a bachelor's degree, 54 completed a master's degree or more, and 20 were unspecified. Their occupations were diverse (e.g., student, retail assistant, doctor, civil servant, paralegal, software developer, digital marketing expert, homemaker etc.) The vast majority had never completed the GRE before, with only 9 having previously taken the exam. Participants' native languages varied; the top three were English (143 participants), Polish (26), and Portuguese (24), with Spanish being a close fourth (19).

7.2 Properties of Learning Examples

Participants were exposed to different learning examples, based on their randomly assigned condition. An overview of the unique properties of the different examples, based on app logs, is reported in Table 2. The Flesch-Kincaid reading ease scores were calculated using the textstat library.[15] The generated sentences were typically longer than the control, while the generated stories were usually the longest. Participants overall spent more time per example for the gen-story condition, followed by the gen-sentence condition, and then the control. On average, the generated stories were the easiest to read, while the generated sentences were the most difficult to comprehend.

7.3 Approaches to Contextualization

In the gen-sentence and gen-story conditions (see Figure 1), participants typed an input to prompt the generation of a custom learning example for a new vocabulary word. The form did not explicitly restrict the length of a user's input, nor did it apply autocompletion or corrections (e.g., for spelling, grammar). We were interested in understanding what types of words and themes people would use to contextualize their learning examples and their corresponding rationales when steering the AI-generations.

7.3.1 ***Which types of user inputs were most commonly used?*** We used the NLTK natural language toolkit [9] to classify users' inputs by word class. As can be seen in Table 3, the majority of inputs were classified as nouns. To give an idea of what people were using to contextualize their examples, we also grouped users' inputs by theme. To do so, we queried ChatGPT for label suggestions

[14] https://aspredicted.org/2p56e.pdf

[15] https://pypi.org/project/textstat/

Table 3: Participants' inputs categorized by word class.

Word Class	Percentage	User Input Examples
Noun	75.7%	animals, video games, friendship
Phrase	15.1%	I love coffee, Convince to give money
Verb / Gerund	4.6%	steal, confront, cooking, gardening, driving
Adjective	2.6%	easy, likeable, arrogant
Proper Noun	1.6%	Taylor Swift, Stardew Valley, Wreck-it Ralph
Other	0.4%	-

for the body of inputs. We then applied cosine similarity between users' inputs and the suggested themes, based on embeddings from OpenAI's embedding model (text-embedding-ada-002). The top five most popular thematic categories across all participants were *Entertainment, Work & Education, Nature, Sports & Health, and Politics.*

7.3.2 ***What were the users' rationales for their inputs?*** Participants in the experimental conditions were asked to explain "*What kinds of things did you try as inputs to generate examples, and why?*" They had different strategies. Based on a thematic analysis, we derived a set of notable themes that had at least 10 mentions (we note the counts in parentheses). Some used their **interests** (22) or things they felt were **relevant to their life experiences** (25). For example, one person wrote: *"I try specific experiences I may have encountered or some things that I interact with every day. It helps me remember the words better given the context."* Others emphasized typing **simple things** in order to yield examples that are more simplistic or easier to understand (17). For instance, one participant wrote that they input *"Generic words like 'food' so it was easier for the software to generate a succinct and clear example to understand the new word in context."* The largest category (72) tried to leverage **associations**; in this case, they tried inputting things that they felt were relevant to the target words or their definitions with the intent to solidify their understanding. For example, one participant stated "*...when I was presented with the word mendacity, I tried to think of situations where a person might possess those attributes and I settled on thief to put the word into a situation that my brain can make a connection...*" Examples of other reasons included gaming the system, humor, etc. The remainder did not specify, or reported no specific rationale such as using random words or the first things that sprang to mind as inputs.

7.4 Impact on Learning Performance

One of our primary pre-registered analyses was to examine whether exposure to the different conditions would influence how well people learn and remember target vocabulary words. To measure this, participants were asked to complete a vocabulary quiz immediately after the learning experience and again one week later. Higher quiz scores indicate greater learning, and a decrease in quiz scores from t1 to t2 would suggest an element of forgetting (or conversely, learning retention). The dependent variable was whether a participant answered a quiz question correctly (correct = 1, incorrect = 0). Since this encoding is discrete and binary, we used a mixed effects logistic regression model where the two timepoints and the three conditions were entered as fixed effects dummy indicator variables. We included random intercepts grouped by participants and vocabulary words to account for multiple sources of non-independence.

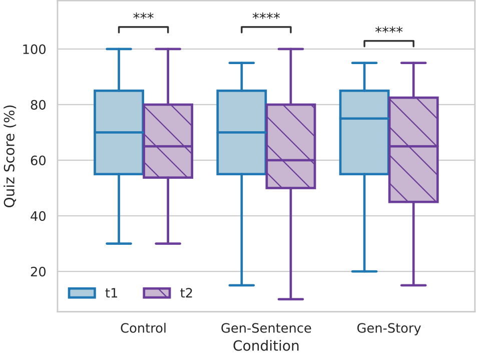


Figure 3: Quiz percentages per condition and timepoint. The statistical annotations shown are based on the Tukey pairwise post hoc tests that were run on the logistic linear regression model. While no differences were found between conditions based on the probability of answering a quiz question correctly, participants were found to score lower on the second quiz compared to the first. (Note: * = $p \leq 0.05$, ** = $p \leq 0.01$, *** = $p \leq 0.001$, **** = $p \leq 0.0001$)

Post hoc pairwise comparisons using Tukey's HSD test detected differences within each condition across timepoints (t1 vs. t2) for the control condition (OR=1.34, 95% CI [1.10, 1.64], $p = 0.0004$), gen-sentence condition (OR=1.54, 95% CI [1.25, 1.89], $p < 0.0001$), and gen-story condition (OR=1.71, 95% CI [1.39, 2.09], $p < 0.0001$). This means that overall scores were observed to be higher overall at t1 over t2 (see Figure 3). However, differences were neither detected between the conditions within each timepoint nor between the conditions for the difference between timepoints. In summary, while we found a difference for the effect of time, there was a lack of evidence that there is a difference between conditions in learning performance.

7.5 Impact on Learning Experience

As another primary pre-registered analysis, we used a mixed effects ordinal logistic regression model to investigate whether the conditions differed by how they impacted people's perception of the learning experience. Learning experience was characterized by the following five subscales: *Interest/Enjoyment, Effort/Importance, Perceived Choice, Value/Usefulness, and Perceived Competence.* The dependent variable was the Likert rating, whereas the three conditions, the five subscales, and their interactions were entered as fixed effects dummy indicator variables. We included random intercepts grouped by participants and question-item to account for multiple sources of non-independence. The results of post hoc pairwise comparisons using the Tukey HSD test are summarized in Figure 4. Overall, the generative AI conditions were both found to be more

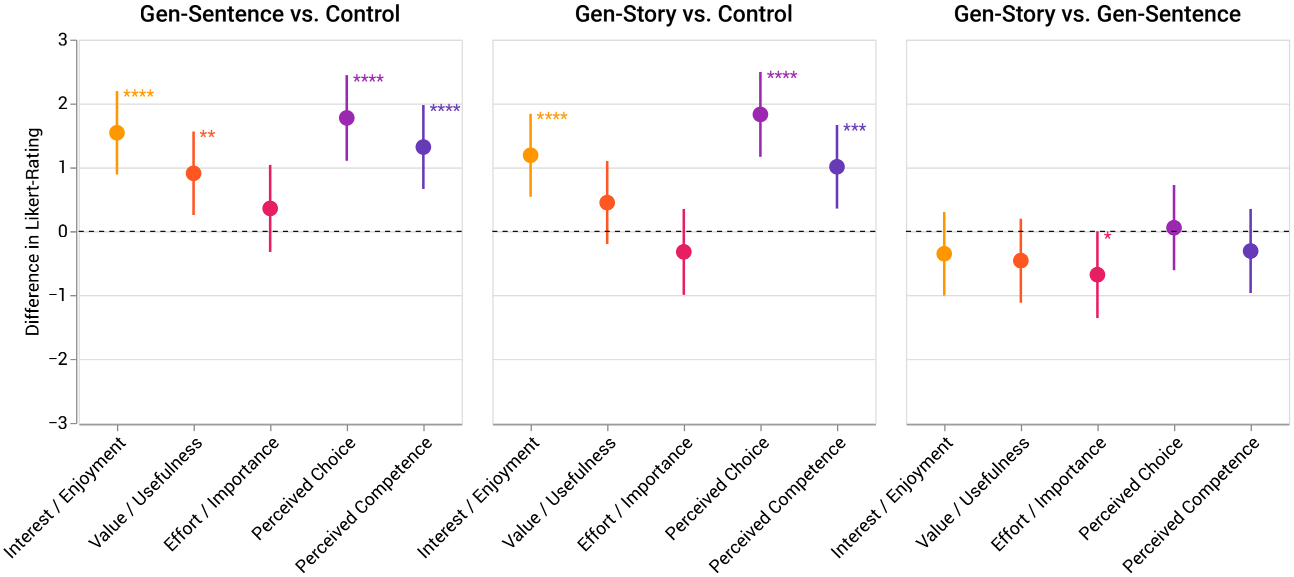


Figure 4: Differences in point estimates with bars depicting the 95% confidence intervals for pairwise comparisons between the conditions along the different IMI subscales. In this comparison, positive values suggest higher values for the condition listed first. (Note: * = p ≤ 0.05, ** = p ≤ 0.01, *** = p ≤ 0.001, **** = p ≤ 0.0001)

intrinsically motivating than the control condition. They also gave participants a greater feeling of choice and competence over the control. However only the gen-sentence condition was found to be more useful than the control.

7.6 Perceived Characteristics of Examples

Participants were asked to rate their overall impression of the learning examples across multiple characteristics. This included whether they found the examples given to them to be *emotionally positive/negative, original/creative, surprising/outrageous, humorous, aligned with their interests, and related to their knowledge.* As an exploratory analysis (not pre-registered), we used a mixed effects ordinal logistic regression model to examine whether exposure to the different conditions influenced people's perception of the examples given to them. Post hoc pairwise comparisons using Tukey's HSD test detected some differences between the conditions based on the perceived qualities, as summarized in Figure 5. Generally, the AI-generated examples aligned with participants by their interests, relatability, and personal relevancy to their knowledge, and were perceived to be more humorous. However, Gen-Sentence was found to be distinctly original/creative compared to the control, and Gen-Story was found to be more emotionally positive overall.

7.7 User Experience

7.7.1 *What attitude did people adopt when using the app?*

Participants rated on a 7-point scale (1= *Very Playfully*, 7 = *Very Seriously*) the attitude they adopted while using the learning application. Overall, participants approached the use of the control condition "*Somewhat Seriously*" (Mdn = 6), the gen-sentence condition "*A Little Seriously*" (Mdn = 5), and the gen-story condition with a "*Neutral*" (Mdn = 4) attitude.

7.7.2 *How much would people want to use the app again or recommend it to others?*

Participants were asked to rate on 7-point scales (1 = *Strongly Disagree*, 7 = *Strongly Agree*) statements regarding whether they would "*...like to use this app again to learn new vocabulary words,*" and "*...recommend this app to a friend to learn new vocabulary words.*" On average, people appraised the gen-sentence condition more highly (Mdn = 6) compared to the control and the gen-story condition (Mdn = 5) for both statements.

7.8 Opportunities & Risks of AI-Generated Learning Content

Participants were asked to write freely about their experience with the vocabulary app. Questions included: "*What did you think about this learning experience,*" "*Is there anything you would want to change or improve about the vocabulary app?*" and "*Any final remarks or comments?*" These questions elicited responses that highlight potential opportunities and pitfalls for AI-generated learning content.

7.8.1 *Opportunities.*

The vast majority of written responses to both the generated-sentence and generated-story conditions were positive. Participants experiencing the generated conditions expressed that they found the experience to be interesting, unique, easy to use and understand, enjoyable, fun, engaging and exciting. Positive quotes included:

- "*It was exciting as **I wasn't expecting the examples to be so fun to read***" (Gen-Sentence)
- "*When I first read the definition I thought of a synonym for that word, then I thought how would I use it in a dialogue, in which context. Then I wrote a keyword regarding such context into the app, and I was actually surprised with the examples it generated; not gonna lie, **I was giggling at how awesome***

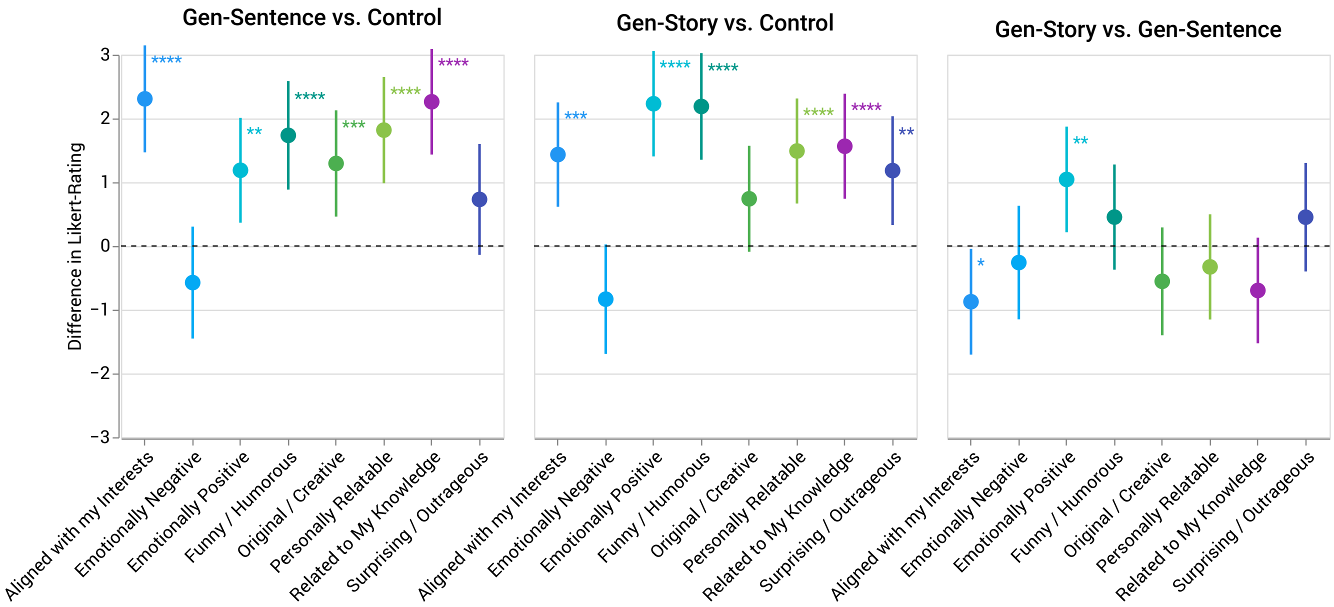


Figure 5: Differences in the point estimates with bars depicting the 95% confidence intervals for pairwise comparisons between the conditions along the different characteristics. In this comparison, positive values suggest higher values for the condition listed first. (Note: * = p ≤ 0.05, ** = p ≤ 0.01, *** = p ≤ 0.001, **** = p ≤ 0.0001)

__this app is__. Then I read the example twice or thrice until I felt I had really understood it." (Gen-Story)

- "*It was easy to understand how it works. It does __require good input from the learner__ - that also makes it more interesting and engaging*" (Gen-Sentence)
- "*It was a different way I have seen to learning new words, I think it is more useful to remember a word when you have to __think about it more critically vs. just memorization__*" (Gen-Sentence)
- "*I think it was a fun experience and its is just a breath of fresh air. And I felt a __deep learning connection__ with the app.*" (Gen-Sentence)
- "*I thought it was a really fun and original way to learn new words, something I have not seen before. The stories were __highly creative__ and I also enjoyed reading them. __I think when you enjoy learning you are more likely to want to stick at it__.*" (Gen-Story)

7.8.2 ***Risks.*** Some participants eluded to possible detractors from such an experience, complaining about some repetitiveness, possible poor-fit or mismatching themes, or the stories being childish. It seems that there is still some challenge in steering AI to offer strong examples in a variety of different ways.

- "*After the first couple of AI generated sentences I could see the usage of the vocab words was __not very natural__. In many occasions, the __definition of the word was placed in the sentence which detracted from focusing__ on how it formed part of the sentence.*" (Gen-Sentence)
- "*The AI generated forced sentences out of __mismatched themes__. This did not show the best examples of usage of the vocab words*" (Gen-Sentence)
- "*It demands a lot of focus as the __examples were pretty long__. I guess learning more than 10 words with it would make me lose my focus and just skip them, making it less effective.*" (Gen-Story)
- "*It felt quite __childish__.*" / "*It could be much better. __The examples were tailored for a kid__.*"(Gen-Story)
- "*It was quite interesting, however after three or four examples I started noticing __patterns in how this AI created its stories and it became boring and predictable__ a tiny bit.*" (Gen-Story)

7.9 Alternative Suggestions

For the participants that experienced the control condition rather than the AI-generated conditions, we noted that some people wished for a multimodal experience, such as being able to hear the pronunciation of a word, or to have more visual stimulation, beyond the text-based interface that was given.

8 DISCUSSION

In this work, we set out to investigate the possibilities and implications of leveraging generative AI for the context personalization of learning materials and conducted an online controlled experiment in the domain of vocabulary learning. We discuss our findings and their implications for future work in this space.

8.1 Characteristics of Generative AI in Supporting Context Personalization

The investigated use of text-to-text generative AI extends beyond prior manual or fill-in-the-blank approaches for context personalization [22, 23, 31, 36, 59] in that it enabled **real-time interactivity** with **deep** levels of personalization. Outputs were more richly connected to a breadth of given topics in unexpected ways, which was appreciated by some users in that they found the examples "exciting," "surprising," and "a breath of fresh air" to read. Comments also alluded to the real-time interaction as encouraging more critical mental engagement with the materials in the moment. However, the investigation also unveiled nuanced insights, which can inform future design considerations for text-based generative AI learning applications. Even with adapting the central theme, series of examples at times exhibited similar structural patterns, which some participants recognized. While we experienced this problem during development and attempted to mitigate it with prompt engineering (for instance by requesting it to refrain from opening with "Once upon a time"), this did not produce enough **structural variety**. Building in mechanisms to slightly vary the underlying prompts may help circumvent this repetition. Generated examples were also occasionally perceived to be somewhat unnatural. This is understandable given the probabilistic nature of generative AI and its flexibility to accept unusual pairings of advanced vocabulary and target input themes. One can speculate that LLMs would be more likely to produce **natural outputs** for less advanced vocabulary that would occur more frequently within the training data. It would therefore be interesting to study how this platform would be perceived when catering to different language learning levels. For HCI and educational technology practitioners, our work offers a practical example for prompting an LLM for the context personalization of vocabulary learning materials, off which others can build and iterate. Furthermore, the reflections distilled from people's reactions towards the generated learning examples in this study can serve as guideposts for similar efforts moving forward.

8.2 Multiple Strategies & Axes for Personalization

We were surprised how people approached context personalization with the app, as indicated in their survey responses (see Section 7.3.2). Besides leveraging words or phrases relatable to their lives and interests to generate examples, a large proportion of participants had other motives. While some strived for simple and easy-to-understand examples to be generated and focused on entering simple inputs, a significant portion of participants attempted to leverage **word-associations** to strengthen their learning, and sought to enter words or phrases that they felt meaningfully related to the target vocabulary word. While this was not anticipated, it highlights a unique practical affordance of generative AI for context personalization. Since users were able to interact directly with the system rather than rely on an educator or researcher as a mediary to facilitate the personalization, the personalization could be driven by users' existing mental models of words and their relationships. Such information was previously infeasible to collect via traditional pre-interviews or surveys. In line with this, it is worth noting that these findings do not exclusively represent context personalization on the basis of interest alone, but speak to user-driven personalization more broadly, mainly in connection to their existing funds of knowledge. This is a useful consideration, with practical implications for the design of generative AI-based learning apps moving forward.

Juxtaposing the positive and critical written feedback with respect to the generative conditions also revealed that besides context personalization, other parameters might need to be customizable to accommodate **different user preferences**. For instance, while some found the stories creative and humorous, others found them repetitive and childish. Furthermore, some participants, but not all, complained that the stories were too long. Taken together, this offers a practical consideration for educators and HCI practitioners aiming to leverage LLMs for context personalization. Personalization based on themes must be considered in concert with other axes of personalization. People can be sensitive to length, tone, and structural repetition within and across multiple textual examples. If these factors are not also catered to with care, the potential benefits of context personalization may be dismissed over frustration with these other aspects. This suggests that an expanded view of personalization that encompasses these attributes may be necessary in future fully fledged app experiences.

8.3 Impact of AI-Enabled Context Personalization

With regards to **learning performance**, the results from this study suggest that AI-enabled user-driven context personalization (i.e., the gen-sentence and gen-story conditions) does not yield *immediate* improvements in learning performance over a non-personalized learning experience (i.e., control condition) in a vocabulary learning task. However, the results in the generated conditions are also not worse than in the control condition, which makes a case for considering the use of generative AI for other positive benefits. Ultimately, this aligned with our expectations since many prior works state that frequency of exposure to vocabulary words affects learners' acquisition of words [17, 42, 51]. Furthermore, it is crucial to note that our study was a short-term exploration that encompassed a one-time exposure to the learning intervention. Therefore, it remains to be seen whether a difference in learning performance would be observed over long-term repeated usage of AI-enabled context personalization for vocabulary learning.

Another area of interest was to determine the impact that AI-enabled context personalization could have on people's **attitudes** towards the learning experience. Levels of **intrinsic motivation**, measured by the Interest/Enjoyment subscale in the IMI, were notably higher in the generative AI learning conditions. This corroborates previous research findings that context personalization increases student motivation [22, 23]. Despite prior literature suggesting that stories boost motivation [39, 40], we could not find evidence that AI-generated stories were more motivating than the AI-generated sentences. Although typing in an interest was a minimal interaction step, the gen-sentence and gen-story conditions provided a substantial increase in their perception of the **choice** they had over their learning experience compared to the control. This indicates that participants perceived the AI-generated content to incorporate their inputs. Nevertheless, it may be worthy to mention that there is a balance to be struck between agency and effort, since some participants complained about the effort needed to type something for each word. One surprising outcome was that although AI-enabled context personalization made our participants feel discernibly more **competent** at learning vocabulary over the control condition, a boost in feelings of competency did not map to higher quiz scores. It would be interesting to investigate the long-term implications of this, since Senko et al. [52] found that an inflated sense of confidence can reduce the time that people would allocate towards studying.

With regards to the **value/usefulness** subscale, we detected that the gen-sentence condition was perceived as more valuable and useful over the control condition, while the story was not. We can speculate that this may be tied to some complaints that the generated stories took longer to read and were repetitive, making them more inclined to ignore them.

In general, we noted that while we did not observe a clear improvement in participants' learning performance from personally contextualized AI-generated examples, their levels of motivation and their overall perception of the learning experience was greatly enhanced. This suggests that a learning process involving continuous practice and repeated exposure may be needed to achieve significant gains, rather than relying on a single session. In light of this, the observed boost in learning motivation attained from the use of AI-enabled context personalization could serve as a potential pathway to sustain regular learning practice, which, in turn, may contribute to improved future learning performance.

While generative AI has to date been speculated as potentially useful in improving education, this work contributes grounding empirical evidence of its utility to foster learning motivation. Additionally, it provides insight that leveraging generative AI in an interactive personalized learning application may also increase learners' perception of autonomy and competency, highlighting potential directions for further investigation. In future, different configurations of variables and different study formats could be considered. For instance, a diary study could be conducted to investigate their influence over an extended period. It could also help to include engagement time as a control variable in upcoming studies since the experimental examples typically took longer for learners to consume and was an aspect that was met with mixed reactions.

8.4 Limitations

Many of our design decisions were motivated by our aim to isolate the effects of AI-enabled context personalization. While striving to make the conditions in the experiment reasonably comparable, we needed to account for multiple learning-example design parameters including UI specifics, targeted length and reading complexity, nuances in the prompt design, etc. Clearly, the number of possible decisions and how they interact with one another forms a high-complexity design space. While it is possible that one may disagree with one or more of these specific design choices, taken together, we observe that they overall helped to shift the needle towards a more rewarding learning experience beyond what is currently available as the state of the art. Besides this, we also note that the study was conducted online. Therefore, we could not enforce that the participants correctly carried out all instructions and did not cheat by referencing external materials. Additionally, despite the rapid growth in popularity of chatGPT and other forms of generative AI, it is possible that some of the participants in this study were unfamiliar with this technology and experienced a novelty effect. Finally, our exploration was a controlled study centered on personally contextualized vocabulary learning material. In an effort to isolate the variables of interest, the intervention was simplistic in its design. As such, more research could be done to understand to what extent these findings transfer to other types of learning scenarios, and how they would play out in real-life contexts. For instance, one could imagine generative AI systems to personalize an entire curriculum to a learner's prior knowledge and interests, guiding them through the learning content with relatable examples and making sure they understand the content before moving on.

8.5 Ethics

In contemplating the broader integration of AI-enabled personalization in learning environments, it is important to consider and cautiously address multiple pertinent ethical issues. We believe that overcoming these challenges will require a collective effort between multiple stakeholders, including educators, policy makers, educational technologists, and HCI researchers.

8.5.1 *Challenges of Using LLMs for Education.*

The adoption of generative AI models such as large language models (LLMs) in education poses significant challenges. LLMs are susceptible to many issues pertaining to toxicity and bias, wherein the models output inappropriate, offensive, or discriminatory content, as well as reliability, wherein the models hallucinate and produce factually inaccurate or nonsensical information [64]. This can impede learners' development, with potentially severe consequences [33]. In our work, we relied on OpenAI's standard guardrails to screen for offensive user inputs, and we leveraged generative AI to output language-focused learning examples in a highly controlled setting, limiting the potential harm to participants. Given that the participants' objective in this study was to expand their vocabulary, and LLMs are language-focused, one can argue that the risks posed by hallucination were relatively limited compared to if the core concept being taught were outside the language-learning domain (e.g., laws of physics). Nevertheless, encountering incongruities regarding the subject matter, even for a language-focused use case, could still distract learners and propagate misinformation if not properly acknowledged.

In the future where AI-enabled personalization is likely to be deployed more broadly, we believe it would be necessary to incorporate additional technological safeguards, such as retrieval-augmented generation and self-checking algorithms [25, 34, 38, 53]. In addition, it would be wise to incorporate the input and supervision of trusted educators and knowledge experts who could vet the quality of the outputs. Until generative AI systems achieve improved accuracy, educators and technological practitioners must critically appraise such content and deploy generative AI judiciously in a manner that supplements rather than replaces established educational resources.

8.5.2 *Data Collection and Privacy.*

The implementation of AI personalization in learning also raises significant data privacy concerns. During personalized content generation, AI models process user inputs potentially containing sensitive or personal information about interests, motivations, or prior knowledge. Service providers employing AI personalization must be transparent about their handling of the data and develop robust policies to prevent data leaks and exploitation of user trust and information. To fully realize the potential of AI-assisted learning, a collaboration between multiple stakeholders is needed to resolve these ethical challenges in order to establish safe, effective, and ethical learning environments.

9 CONCLUSION

In light of recent enthusiasm for generative AI to revolutionize education, we take a pause to critically examine its potential to improve learning outcomes and attitudes through deep and dynamic context personalization. Context personalization involves adapting learning materials to be grounded in users' interests, preferences, and existing funds of knowledge. In this paper, we implemented a custom vocabulary app that facilitated three distinct learning conditions: a control condition in which examples of how words can be used in a sentence were based off of pre-existing texts, and two experimental conditions in which examples were dynamically generated with AI (gpt-3.5-turbo). Based off a between-subjects experiment with 272 participants that involved a lesson, survey, and two quizzes, we gained several insights. In this paper, we contributed grounding empirical evidence of its utility to boost learning motivation, as well as several research and design considerations specific to this use case, on the basis of people's varied preferences and strategies for contextualizing their own learning examples. Ultimately, we acquired the perspective that while one-shot, short-term exposure to AI-generated personalized learning examples does not result in immediate improvements in learning performance, it can boost people's feelings of intrinsic motivation in learning. As such, AI may indeed be able to cheaply and scalably afford rich individualized adaptations of learning examples, and we speculate that this may lead to improvements in learning performance over the long term.

ACKNOWLEDGMENTS

Statistical support was provided by data science specialist Steven Worthington, at the Institute for Quantitative Social Science, Harvard University. We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), [PGS D3-545858-2020]. Cette recherche a été financée par le Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG), [PGS D3-545858-2020]. We also thank the MIT Integrated Learning Initiative for helping fund this project.

REFERENCES

[1] 2005. Qualtrics. https://www.qualtrics.com Accessed: 2023-11-17.

[2] 2014. Intrinsic Motivation Inventory (IMI). https://selfdeterminationtheory.org/intrinsic-motivation-inventory Accessed: 2023-11-17.

[3] 2014. Prolific. https://www.prolific.com Accessed: 2023-11-17.

[4] Mary Ainley, Suzanne Hidi, and Dagmar Berndorff. 2002. Interest, learning, and the psychological processes that mediate their relationship. *Journal of educational psychology* 94, 3 (2002), 545.

[5] Hamdan Alamri, Victoria Lowell, William Watson, and Sunnie Lee Watson. 2020. Using personalized learning as an instructional approach to motivate learners in online higher education: Learner self-determination and intrinsic motivation. *Journal of Research on Technology in Education* 52, 3 (2020), 322–352.

[6] Sultan Altalhab. 2018. Short-and long-term effects of repetition strategies on vocabulary retention. *Advances in Language and Literary Studies* 9, 2 (2018), 146–149.

[7] Riku Arakawa, Hiromu Yakura, and Sosuke Kobayashi. 2022. VocabEncounter: NMT-powered Vocabulary Learning by Presenting Computer-Generated Usages of Foreign Words into Users' Daily Lives. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–21.

[8] Sasha Barab, Tyler Dodge, Hakan Tuzun, Kirk Job-Sluder, Craig Jackson, Anna Arici, Laura Job-Sluder, Robert Carteaux Jr, Jo Gilbertson, and Conan Heiselt. 2007. The Quest Atlantis Project: A socially-responsive play space for learning. In *The design and use of simulation computer games in education*. Brill, 159–186.

[9] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

[10] Dale Brown. 2008. Using a modified version of the Vocabulary Knowledge Scale to aid vocabulary development. *The Language Teacher* 32, 12 (2008), 15–16.

[11] Tanya L Chartrand and John A Bargh. 1996. Automatic activation of impression formation and memorization goals: Nonconscious goal priming reproduces effects of explicit task instructions. *Journal of personality and Social Psychology* 71, 3 (1996), 464.

[12] Fiona Draxler, Laura Haller, Albrecht Schmidt, and Lewis L Chuang. 2022. Auto-Generating Multimedia Language Learning Material for Children with Off-the-Shelf AI. In *Proceedings of Mensch und Computer 2022*. 96–105.

[13] Fiona Draxler, Audrey Labrie, Albrecht Schmidt, and Lewis L Chuang. 2020. Augmented reality to enable users in learning case grammar from their real-world interactions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[14] Fiona Draxler, Albrecht Schmidt, and Lewis L Chuang. 2023. Relevance, Effort, and Perceived Quality: Language Learners' Experiences with AI-Generated Contextually Personalized Learning Material. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*. 2249–2262.

[15] Barbara L Fredrickson. 2004. The broaden–and–build theory of positive emotions. *Philosophical transactions of the royal society of London. Series B: Biological Sciences* 359, 1449 (2004), 1367–1377.

[16] Anne C Frenzel, Reinhard Pekrun, Anna-Lena Dicke, and Thomas Goetz. 2012. Beyond quantitative decline: conceptual shifts in adolescents' development of interest in mathematics. *Developmental psychology* 48, 4 (2012), 1069.

[17] Yongqi Gu and Robert Keith Johnson. 1996. Vocabulary learning strategies and language learning outcomes. *Language learning* 46, 4 (1996), 643–679.

[18] David L Hamilton, Lawrence B Katz, and Von O Leirer. 1980. Cognitive representation of personality impressions: Organizational processes in first impression formation. *Journal of Personality and Social Psychology* 39, 6 (1980), 1050.

[19] Ellen Karoline Henriksen, Justin Dillon, and Jim Ryder. 2015. *Understanding student participation and choice in science and technology education*. Springer.

[20] Suzanne Hidi and Judith M Harackiewicz. 2000. Motivating the academically unmotivated: A critical issue for the 21st century. *Review of educational research* 70, 2 (2000), 151–179.

[21] Suzanne Hidi and K Ann Renninger. 2006. The four-phase model of interest development. *Educational psychologist* 41, 2 (2006), 111–127.

[22] Sigve Høgheim and Rolf Reber. 2015. Supporting interest of middle school students in mathematics through context personalization and example choice. *Contemporary Educational Psychology* 42 (2015), 17–25.

[23] Sigve Høgheim and Rolf Reber. 2017. Eliciting mathematics interest: New directions for context personalization and example choice. *The Journal of Experimental Education* 85, 4 (2017), 597–613.

[24] Adam Ibrahim, Brandon Huynh, Jonathan Downey, Tobias Höllerer, Dorothy Chun, and John O'donovan. 2018. Arbis pictus: A study of vocabulary learning with augmented reality. *IEEE transactions on visualization and computer graphics* 24, 11 (2018), 2867–2874.

[25] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *Comput. Surveys* 55, 12 (2023), 1–38.

[26] J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. (1975).

[27] Ronnel B King, Dennis M McInerney, Fraide A Ganotice Jr, and Jonalyn B Villarosa. 2015. Positive affect catalyzes academic engagement: Cross-sectional, longitudinal, and experimental evidence. *Learning and individual differences* 39 (2015), 64–72.

[28] Stanley B Klein and Judith Loftus. 1990. Rethinking the role of organization in person memory: An independent trace storage model. *Journal of Personality and Social Psychology* 59, 3 (1990), 400.

[29] Geza Kovacs and Robert C Miller. 2014. Smart subtitles for vocabulary learning. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 853–862.

[30] Andreas Krapp. 2002. Structural and dynamic aspects of interest development: Theoretical considerations from an ontogenetic perspective. *Learning and instruction* 12, 4 (2002), 383–409.

[31] Heng-Yu Ku, Christi A Harter, Pei-Lin Liu, Ling Thompson, and Yi-Chia Cheng. 2007. The effects of individually personalized computer-based instructional program on solving mathematics problems. *Computers in human behavior* 23, 3 (2007), 1195–1210.

[32] Joanne Leong. 2023. Using Generative AI to Cultivate Positive Emotions and Mindsets for Self-Development and Learning. *XRDS: Crossroads, The ACM Magazine for Students* 29, 3 (2023), 52–56.

[33] Stephan Lewandowsky, Ullrich KH Ecker, Colleen M Seifert, Norbert Schwarz, and John Cook. 2012. Misinformation and its correction: Continued influence and successful debiasing. *Psychological science in the public interest* 13, 3 (2012), 106–131.

[34] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474.

[35] Matthew D Lieberman. 2012. Education and the social brain. *Trends in Neuroscience and Education* 1, 1 (2012), 3–9.

[36] Cecilia L López and Howard J Sullivan. 1992. Effect of personalization of instructional context on the achievement and attitudes of Hispanic students. *Educational Technology Research and Development* 40, 4 (1992), 5–14.

[37] Mircea F Lungu, Luc van den Brand, Dan Chirtoaca, and Martin Avagyan. 2018. As we may study: Towards the web as a personalized language textbook. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–12.

[38] Potsawee Manakul, Adian Liusie, and Mark JF Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896* (2023).

[39] Scott W McQuiggan, Jonathan P Rowe, Sunyoung Lee, and James C Lester. 2008. Story-based learning: The impact of narrative on learning experiences and outcomes. In *Intelligent Tutoring Systems: 9th International Conference, ITS 2008, Montreal, Canada, June 23-27, 2008 Proceedings 9*. Springer, 530–539.

[40] Mai Murmann and Lucy Avraamidou. 2014. Animals, emperors, senses: Exploring a story-based learning design in a museum setting. *International Journal of Science Education, Part B* 4, 1 (2014), 66–91.

[41] Rebecca L Oxford and Robin C Scarcella. 1994. Second language vocabulary learning among adults: State of the art in vocabulary instruction. *System* 22, 2 (1994), 231–243.

[42] Rolf Palmberg. 1987. Patterns of Vocabulary Development in Foreign-Language Learners1. *Studies in second language acquisition* 9, 2 (1987), 201–219.

[43] T Sima Paribakht and Marjorie Bingham Wesche. 1993. Reading comprehension and second language development in a comprehension-based ESL program. *TESL Canada journal* (1993), 09–29.

[44] Pat Pataranutaporn, Joanne Leong, Valdemar Danry, Alyssa P Lawson, Pattie Maes, and Misha Sra. 2022. AI-Generated Virtual Instructors Based on Liked or Admired People Can Improve Motivation and Foster Positive Emotions for Learning. In *2022 IEEE Frontiers in Education Conference (FIE)*. IEEE, 1–9.

[45] Zhenhui Peng, Xingbo Wang, Qiushi Han, Junkai Zhu, Xiaojuan Ma, and Huamin Qu. 2023. Storyfier: Exploring Vocabulary Learning Support with Text Generation Models. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–16.

[46] Junaid Qadir. 2023. Engineering education in the era of ChatGPT: Promise and pitfalls of generative AI for education. In *2023 IEEE Global Engineering Education Conference (EDUCON)*. IEEE, 1–9.

[47] Rolf Reber, Elizabeth A Canning, and Judith M Harackiewicz. 2018. Personalized education to increase interest. *Current directions in psychological science* 27, 6 (2018), 449–454.

[48] K Ann Renninger and Suzanne E Hidi. 2015. *The power of interest for motivation and engagement.* Routledge.

[49] Henry L Roediger III and Jeffrey D Karpicke. 2006. Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological science* 17, 3 (2006), 249–255.

[50] Richard M Ryan. 1982. Control and information in the intrapersonal sphere: An extension of cognitive evaluation theory. *Journal of personality and social psychology* 43, 3 (1982), 450.

[51] Thomas Saragi et al. 1978. Vocabulary learning and reading. *System* 6, 2 (1978), 72–8.

[52] Corwin Senko, Andrew H Perry, and Melissa Greiser. 2022. Does triggering learners' interest make them overconfident? *Journal of Educational Psychology* 114, 3 (2022), 482.

[53] Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567* (2021).

[54] U.S. Department of Education. 2017. Reimagining the Role of Technology in Education: 2017 National Education Technology Plan Update. (2017). https://tech.ed.gov/files/2017/01/NETP17.pdf

[55] Christian Vázquez, Lei Xia, Takako Aikawa, and Pattie Maes. 2018. Words in motion: Kinesthetic language learning in virtual reality. In *2018 IEEE 18th International Conference on advanced learning technologies (ICALT)*. IEEE, 272–276.

[56] Christian David Vazquez, Afika Ayanda Nyati, Alexander Luh, Megan Fu, Takako Aikawa, and Pattie Maes. 2017. Serendipitous language learning in mixed reality. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. 2172–2179.

[57] Candace Walkington and Matthew L Bernacki. 2014. Motivating students by "personalizing" learning around individual interests: A consideration of theory, design, and implementation issues. In *Motivational interventions*. Emerald Group Publishing Limited.

[58] Candace Walkington and Matthew L Bernacki. 2018. Personalization of instruction: Design dimensions and implications for cognition. *The Journal of Experimental Education* 86, 1 (2018), 50–68.

[59] Candace A Walkington. 2013. Using adaptive learning technologies to personalize instruction to student interests: The impact of relevant contexts on performance and learning outcomes. *Journal of educational psychology* 105, 4 (2013), 932.

[60] Melissa B Wanzer, Ann B Frymier, and Jeffrey Irwin. 2010. An explanation of the relationship between instructor humor and student learning: Instructional humor processing theory. *Communication education* 59, 1 (2010), 1–18.

[61] Maheshya Weerasinghe, Verena Biener, Jens Grubert, Aaron Quigley, Alice Toniolo, Klen Čopič Pucihar, and Matjaž Kljun. 2022. Vocabulary: Learning vocabulary in ar supported by keyword visualisations. *IEEE Transactions on Visualization and Computer Graphics* 28, 11 (2022), 3748–3758.

[62] Ben Williamson. 2009. *Computer games, schools, and young people: A report for educators on using games for learning.* Futurelab Bristol.

[63] 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. (2022).

[64] Terry Yue Zhuo, Yujin Huang, Chunyang Chen, and Zhenchang Xing. 2023. Exploring ai ethics of chatgpt: A diagnostic analysis. *arXiv preprint arXiv:2301.12867* (2023).