

# Storyfier: Exploring Vocabulary Learning Support with Text Generation Models

Zhenhui Peng\*  
pengzhh29@mail.sysu.edu.cn  
Sun Yat-sen University  
Zhuhai, China

Xingbo Wang\*  
xwangeg@cse.ust.hk  
The Hong Kong University of Science  
and Technology  
Hong Kong, China

Qiushi Han  
Sun Yat-sen University  
Zhuhai, China  
hanqsh@mail2.sysu.edu.cn

Junkai Zhu  
Guangdong Polytechnic of Industry &  
Commerce  
Guangzhou, China  
zhujunkai@hotmail.com

Xiaojuan Ma  
The Hong Kong University of Science  
and Technology  
Hong Kong, China  
mxj@cse.ust.hk

Huamin Qu  
The Hong Kong University of Science  
and Technology  
Hong Kong, China  
huamin@cse.ust.hk

## ABSTRACT

Vocabulary learning support tools have widely exploited existing materials, e.g., stories or video clips, as contexts to help users memorize each target word. However, these tools could not provide a coherent context for any target words of learners' interests, and they seldom help practice word usage. In this paper, we work with teachers and students to iteratively develop Storyfier, which leverages text generation models to enable learners to read a generated story that covers any target words, conduct a story cloze test, and use these words to write a new story with adaptive AI assistance. Our within-subjects study (N=28) shows that learners generally favor the generated stories for connecting target words and writing assistance for easing their learning workload. However, in the read-cloze-write learning sessions, participants using Storyfier perform worse in recalling and using target words than learning with a baseline tool without our AI features. We discuss insights into supporting learning tasks with generative models.

## CCS CONCEPTS

- **Human-centered computing** → **User interface design**; • **Computing methodologies** → **Natural language generation**; • **Applied computing** → *Computer-assisted instruction*.

## KEYWORDS

vocabulary learning, story generation, language models

## ACM Reference Format:

Zhenhui Peng, Xingbo Wang, Qiushi Han, Junkai Zhu, Xiaojuan Ma, and Huamin Qu. 2023. Storyfier: Exploring Vocabulary Learning Support with Text Generation Models. In *The 36th Annual ACM Symposium on User Interface*

\*Both authors contributed equally to this research.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).  
UIST '23, October 29–November 1, 2023, San Francisco, CA, USA  
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.  
ACM ISBN 979-8-4007-0132-0/23/10...\$15.00  
<https://doi.org/10.1145/3586183.3606786>

*Software and Technology (UIST '23), October 29–November 1, 2023, San Francisco, CA, USA. ACM, New York, NY, USA, 16 pages. <https://doi.org/10.1145/3586183.3606786>*

## 1 INTRODUCTION

Learning vocabulary in meaningful contexts, such as stories and images in language learning textbooks, and video clips from movies, is a common and effective practice as it enables deep and active processing of vocabulary (e.g., word associations, logic) [54]. Many existing vocabulary learning systems like VocabEncounter [1] and Smart Subtitles [36] have exploited a variety of materials to establish the contexts for words. These systems have demonstrated that the provided contexts can enhance vocabulary memorization [1, 36].

However, these systems may fall short in two aspects. First, they largely leverage existing materials and could not provide a meaningful context for any set of target words that users wish to learn. In other words, previous systems lack the flexibility to offer a story, an article, or a video clip that covers the target words that teachers or learners specify. These coherent contexts that connect target words may make a difference to vocabulary learners. As suggested by Gu *et al.* [25], learning vocabulary in batches under coherent contexts could facilitate recalls of a larger amount of words compared to learning vocabulary in isolation. Second, previous systems primarily focus on helping users to understand and memorize the meanings of target words via meaning-focused input learning activities, e.g., reading and listening, that use language receptively [48]. Few systems facilitate learners to master the usage of learned words via productive and fluency development tasks (e.g., writing and speaking) – typical activities that could help master the meanings and usage of target words in traditional courses [48]. In offline courses, teachers can provide in-situ adaptive support like hints on word usage during these learning activities; however, this is often unavailable to individual learners outside classrooms.

In this work, we utilize stories generated by large language models (LLMs) as meaningful contexts that cover any target words and provide adaptive assistance in word usage practice. Our focus is motivated, on one hand, by the prevalent use of stories in language learning textbooks, and the proven efficacy of story-based learning in various scenarios such as programming [13, 14, 73], parent-child storytelling [92], and children's visual storytelling [91]. On the

other hand, LLMs can generate fluent and relevant texts given user specifications such as keywords, which have been used to support the writings of emails [24], articles or fiction [10], and poems [71, 78]. However, little work, if any, has explored LLMs for story-based vocabulary learning where users should spend effort in mastering target words' meanings and usage. Questions arise such as 1) whether and how LLMs can generate the meaningful context of any target word set for vocabulary learning, 2) if so, what vocabulary learning activities can these generative models support, and 3) how would the support from generative models impact the users' vocabulary learning outcome and experience.

To this end, we seek to provide insights into these questions by designing, developing, and evaluating an AI-generated story-based vocabulary learning system, *Storyfier*, that can provide meaningful story contexts and adaptive assistance for learning any set of target words. Here, we choose English as the target language to learn and target ESL (English-as-the-Second-Language) Chinese learners, e.g., high-school or university students in China. We take an iterative design approach with insights from educational literature and the involvement of teachers, learners, and HCI researchers in this process. We first fine-tune a text-generation model on a short-story corpus and validate its capability in producing meaningful story context given a set of target CET-4 English words<sup>1</sup>. We then present this model to three English teachers and five experienced ESL learners in an interview study to explore possible learning activities that *Storyfier* can support. Based on the insights from the interviews and educational literature, we develop a *Storyfier* prototype that supports three types of vocabulary learning activities: 1) *reading* an AI-generated story with target words, 2) solving story *cloze* tests on target words (i.e., fill blanks of the generated story by using target words), and 3) *writing* a story using target words with the AI models by turns. We seek feedback on *Storyfier*'s design and refine it via a user study with twelve ESL learners and two co-design workshops with the three English teachers mentioned above and four HCI researchers.

We conduct a  $2 \times 2$  within-subjects study with 28 university students to evaluate the impact of *Storyfier*'s AI functions (with vs. without generative models) and learning activity (read-only vs. read-cloze-write) on the learning outcome and experience. The results show that in the read-only learning sessions, the generative stories do not help to improve learning gains in recalling target words' meanings and mastering their usage. In read-cloze-write learning sessions, participants with generated stories and AI assistance perform even worse compared to the condition without the generative models. However, most participants still indicate their preferences on *Storyfier*'s generated stories for connecting target words and its writing assistance for reducing learning workload. Based on our findings, we highlight the value of generative models in offering meaningful materials and enjoyable experience for learning tasks. We also urge future AI-supported learning tools to ensure users to spend the necessary effort in their learning tasks.

Our work makes three contributions. First, we present a vocabulary learning system *Storyfier* that facilitates users to master the meanings and usage of any target English words via AI-generated

stories and writing assistance. Second, our design and evaluation of *Storyfier* provide first-hand findings on the feasibility, effectiveness, and user experience of applying generative models to vocabulary learning. Third, we offer insights and design considerations of leveraging generative models to support learning tasks.

## 2 RELATED WORK

To situate our work, we start by reviewing the pedagogical strategies and activities for vocabulary learning. We then discuss previous vocabulary learning support systems. Lastly, we introduce related textual story-generation techniques that enable us to achieve the envisioned *Storyfier*.

### 2.1 Pedagogical Strategies and Activities For Vocabulary Learning

According to the amount of context information used, vocabulary learning strategies can be categorized as decontextualized, partially-contextualized, and fully-contextualized [26, 52]. Decontextualized techniques, including using word lists in alphabetical order or by part of speech, flashcards, dictionary, focus on learning isolated words without meaningful contexts. For example, *dictionary* provides detailed instructions on grammar, pronunciation, and brief usage examples. However, improper use of dictionary, e.g., checking every word's meaning during reading and failing to associate it with the current context, would result in poor learning outcomes [74]. In other words, decontextualized techniques may not aid long-term vocabulary retention and practical word usage [25, 75].

Educators have argued that vocabulary is better learned through contextualized learning activities [26]. Partially-contextualized techniques provide a certain amount of context information (e.g., word association). For instance, *Word grouping* organizes words according to different criteria, such as (dis)similarity and topic. *Concept association* (or "elaboration") constructs connections between new words and some familiar contexts, such as previously learned words, personal experience, or knowledge in learners' memory [7]. Besides, *keyword techniques* [59] link words with visual [2] or aural [16] objects to improve vocabulary memorization. **Fully contextualized techniques** associate words in fully authentic communication contexts and connect them with a meaningful flow (e.g., logic), which are considered the peak of L2 vocabulary learning techniques [52]. They use existing newspapers, articles, magazines, and novels as learning material. The most common activities are reading or listening to the stories in contextual inference tasks (e.g., cloze test) [26]. Speaking and writing practices are regarded as the more effective but also challenging activities, which require turning receptive vocabulary knowledge into productive use in communication contexts [53]. Nevertheless, contextualized methods are demanding and complex for individual learners and are usually adopted by teachers in classroom activities [26, 69].

Regarding the learning activities in a traditional language course, Paul Nation, suggested that there should be roughly equal amounts of time given to each of the following four strands [47]. The **meaning-focused input** strand involves learning through listening and reading – using language receptively. This strand mainly focuses on understanding what they listen to and read, e.g., stories, TVs, films, conversations, and so on. The **meaning-focused output** strand

<sup>1</sup>Short for College English Test Band 4, a mandatory test for acquiring bachelor degrees in China.

involves learning through speaking and writing – using language productively. Typical activities in this strand include talking in conversations, writing a letter or a note, keeping a diary, telling a story, etc. The **language-focused learning** strand involves the deliberate learning of language features such as pronunciation, spelling, vocabulary, grammar, and discourse. Lastly, the **fluency development strand** should involve four skills of listening, speaking, reading, and writing. In this strand, learners are helped to make the best use of what they have already known in typical activities like ten-minutes writing and listening to easy stories. These four strands can fit together in many different ways [47, 48]. For example, a group collaborative writing activity in the high-school can combine the meaning-focused output and language-focused learning strands, if the output written work deliberately focuses on the vocabulary and grammar [47].

Our work is motivated by the benefits of fully contextualized strategies and gets inspired by the four strands of activities for vocabulary learning. We use textual short stories as contextualized vocabulary learning materials. We support individual vocabulary learners with a proper integration of the four strands of learning activities based on the story contexts.

## 2.2 Vocabulary Learning Support Systems

Researchers have proposed various approaches and systems to support vocabulary learning. For rote learning, a bunch of work manage to model users’ memory cycles and plan the target words with proper difficulty level and repetition frequency [8, 49, 50, 90]. As for our focused contextual learning, previous vocabulary learning systems have exploited materials in different mediums, such as images [76], physical locations [88], textual articles in webpage [1], subtitles of videos [32, 36, 67], and augmented/virtual reality [31, 60, 68]. For instance, FinDo [88] is a mobile application that helps users understand the vocabulary about the surrounding objects with the contexts of users’ current locations. Tangworakitthaworn *et al.* [76] used image processing techniques to extract visual objects in photos and matched them with the target vocabulary. VocabEncounter [1] encloses target vocabulary into reading materials to facilitate micro learning in daily life. Smart Subtitles [36] equips video subtitles with features like vocabulary definitions on hover and dialog-based video navigation.

However, these systems largely make use of existing materials as contexts, which may not be able to provide a meaningful flow that covers any set of target words – a requirement of fully contextualized learning [52]. We seek to mitigate this constraint by generating a short story for any target word set. Our decision to use stories as the context for words is inspired by their common usage in language learning textbooks and the proven efficacy of story-based learning in other scenarios [13, 14, 73, 91, 92]. Further, previous vocabulary learning support systems mainly focus on supporting meaning-focused input activities that aim at understanding the words’ meanings. Our work further supports other types of learning activities that help to master the usage of target words.

## 2.3 Textual Story Generation Techniques

Recent advances in textual story generation offer potentials to support vocabulary learners with meaningful contexts that cover

any word set and offer in-situ learning support. The textual story generation techniques aim at generating coherent and fluent narratives or ideas based on simple user inputs, such as a title [46] and prompts [17]. Early computational work adopts symbolic approaches [58, 64, 84] that first select a sequence of characters and actions according to aesthetic, narrative conflicts, and logic, and then create a story with pre-defined templates. Another approach is case-based reasoning [19, 77], which extracts the story plots of existing stories and adapts them to new contexts. Yet, these methods are restricted by predefined story domains and styles. Recent story generation methods mainly adopt sequence-to-sequence language models [17, 27, 44, 89], which can learn complex and implicit relationships among story plots. Particularly, transformer-based models [4, 35, 70] are able to produce incredibly fluent texts after training on a large language corpus. These models can be finetuned to support downstream applications like writing assistants [5, 6] and health consultation [82].

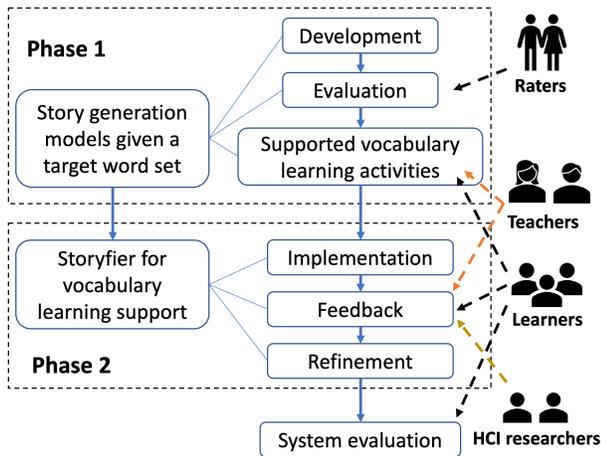
To generate stories with desired properties (*e.g.*, keywords, topic, styles), researchers apply techniques like decoding strategies, prompt controls, and finetuning to build controllable language models. Decoding strategies aim to restrict and influence the sampling process of generation to change the features of output texts. These features can describe the user preferences and are modeled by heuristics [20], supervised signals [30], and reinforcement learning [40]. Prompt controls use natural language (*e.g.*, “translate to English”) to elicit desired contents [4, 33, 39, 41, 43, 61, 72]. Finetuning methods investigate effective conditional training based on key words [17], story valence [55], character fortune [9], control codes [35] (*e.g.*, , topic, sentiment), and simpler attribute models [11, 37].

Recent intelligent systems have explored the usage of text generation techniques in a variety of scenarios, such as creative writing [45, 57], AI-mediated communication [18], and health intervention [34]. In the story-based learning scenario, StoryBuddy [92] assists parents-children storytelling via a question-answer generation model, which consists of a rule-based answer generation module, a BART-based question generation module, and a ranking module. It can help parents create a storytelling bot that can tell stories, ask children questions, and provide feedback [92]. However, these studies present a different focus compared to ours. We specifically investigate the use of and interaction with text generation models for story-based vocabulary learning.

In this paper, we first customize and evaluate a controllable language model for generating meaningful stories that cover given target word set. We then explore what vocabulary learning activities that this model can support with teachers and students.

## 3 PHASE 1: FEASIBILITY AND SUPPORTED ACTIVITIES OF STORY GENERATION MODELS FOR VOCABULARY LEARNING

To help individual learners to master the meaning and usage of any target word sets, we design and develop *Storyfier* via a two-phase process (Figure 1). In this section, we present the first phase in which we 1) develop a story generation model, 2) validate its feasibility for providing meaningful contexts for vocabulary learning, and 3) explore what vocabulary learning activities this model could support.



**Figure 1: Our two-phases design and development process of *Storyfier* with teachers, learners, and HCI researchers.**

### 3.1 Developing Story Generation Models

Given the potential benefits of a meaningful story for learning a batch of words [25], we first seek to develop a controllable language model that can generate stories with given target word sets. Here, we target the vocabulary pool (4,827 in total) required by the College English Test Band 4 (CET-4), a mandatory national test for Chinese university students to obtain bachelor degrees.

**3.1.1 Dataset.** We choose ROCStory corpus [46] to contextualize CET-4 words and build story generation models. ROCStory collects over 100,000 five-sentence commonsense human-written stories (Table 1<sup>2</sup>). The simple and short story form could help learners easily understand the story flow and mitigate diversion from vocabulary learning to story comprehension. The simplicity of the story structures and logic is also appreciated by English teachers who participate in the later studies (subsection 3.3.2). Though these stories are short, they are created by various human workers and have passed qualification tests to ensure story quality and creativity. In addition, these stories have causal and temporal commonsense relationships between story sentences and cover a wide range of everyday topics, such as movie, school, birthday, and music. Therefore, if there is a story that covers a set of target words, learners can easily associate a group of words with a common topic following a meaningful logic flow. With this dataset, we aim to develop a model that can generate meaningful stories like those in ROCStory given any set of target words in the CET-4 pool.

**3.1.2 Data Preprocessing and Model Building.** Figure 2 summarizes our data preprocessing and model building procedure. Specifically, we follow recent story generation techniques [9, 17, 55] and formulate the problem as a sequence-to-sequence translation task. We first segment the stories into titles<sup>3</sup> and sentences. Then, using the CET-4 word list, we identify the occurrences of these words in each sentence of every story and sort them chronologically. This leads to the creation of a set of {*story title*, *target words*, *story sentences*}

<sup>2</sup>Readability is measured by Flesch Reading-Ease, and CET-4 is 34.23 on average.

<sup>3</sup>For the stories without titles, we represent their title features as "no title".

**Table 1: The statistics of ROCStory dataset.**

Attributes	Values
# of stories	101,661
# of words	4,640,319
Average story length	45.65
Average sentence length	7.80
Average readability	57.14
Coverage of CET-4 words	89.52%

tuples (Figure 2A). For story generation, we leverage a state-of-the-art open-source language model T5 [62] as the base model. Our decision is made based on two reasons. First, T5 exhibits impressive performance across various NLP tasks (e.g., text generation and classification), which can be attributed to its unified text-to-text framework and its pretraining on a large language corpus. Moreover, it is freely available and adaptable to our application scenario compared to other impressive but closed-source language models (e.g., GPT-3 [4] and GPT-4 [51]).

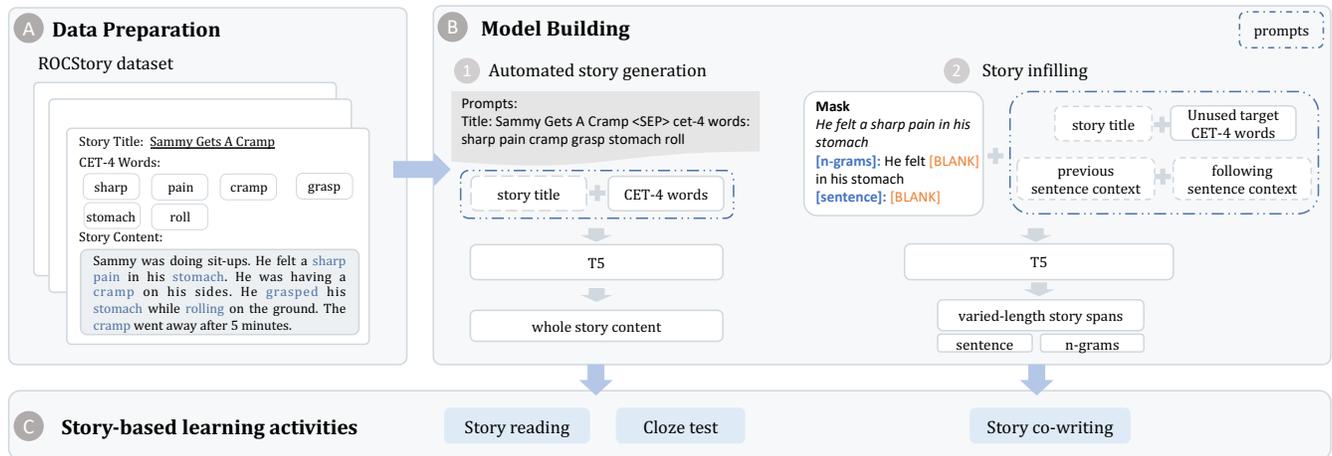
Then, we adopt a prompt-based approach to finetune and steer the model generation process to learn the mappings between target English words and a story (Figure 2B1). We formulate the input prompt as the concatenation of the *story title* and *target words* derived from story tuples of the processed dataset. According to our experiments, the *title* imposes a high-level control of story relevance and leads to faster and better convergence compared to training without the *title* signal. We finetune the pretrained T5-large model offered in the HuggingFace on our dataset using Adam optimization algorithm with a 0.0001 learning rate. The training process lasts for five epochs and has a 0.9857 cross-entropy loss.

After training, our model can generate complete stories rather than isolated sentences, thus creating meaningful contexts for the target words across multiple sentences. For instance, when given the words “athlete”, “avid”, and “frequently” (as shown in Figure 3), the model begins a narrative about an avid athlete who frequently participates in marathons.

### 3.2 Evaluating the Quality of Generative Stories

While our model can generate a story given any word set with or without a title, at this stage, we would like to compare the quality of machine-generated and human-written stories in the corpus which cover the same target word set. This evaluation aims at validating if the generated stories were competent for vocabulary learning support. We will assess the perceived quality and helpfulness of generated stories given any target words in our later interviews with teachers (subsection 3.3.2) and experiments with learners (subsection 6.3.2). Following prior work [9, 17, 55], we conduct technical and human evaluations. We sample 20 stories from the ROCStory dataset with varied difficulty levels (i.e., word frequency) of contained CET-4 words. For each human-written story, we use our trained language models to generate a machine version based on the story title and contained CET-4 words. Thereafter, we create 20 human-machine story pairs (40 stories in total).

**3.2.1 Technical Evaluation.** We assess the story content from grammatical accuracy, lexical diversity (i.e., number of unique words, and



**Figure 2: The technical framework of *Storyfier*.** We mainly adopt prompt-based fine-tuning strategies to build story generation models. (A) We derive CET-4 words from the stories in ROCStory dataset. (B) We finetune T5 language models to 1) generate a story given a CET-4 word set with or without a title (presented in section 3.2) and 2) infill a sentence or n-grams given preceding and following sentences, unused target words, and story title (if any) (subsection 4.2.1). (C) We apply the models to support three kinds of story-based learning activities.

**Table 2: Automated evaluation of human-written and machine-generated stories using lexical metrics.**

	Grammar	Type-token ratio	Trigram repetition	Sentence coherence
Human	1.00	0.75	0.01	0.42
Machine	1.00	0.77	0.01	0.43

**Table 3: Average human ratings of machine-generated and human-written stories. (\*:  $p < .05$  using Wilcoxon Signed-rank test)**

	Coherence *	Relevance *	Interestingness	Overall
Human	4.53	4.58	4.08	4.26
Machine	3.92	4.33	3.97	4.01

type-token ratio: number of unique words/total number of words), and lexical coherence (*i.e.*, trigram repetition, and sentence coherence<sup>4</sup>: average semantic similarities between sentences) [22, 38, 65]. As shown in Table 2, both the machine-generated and human-written stories have no grammar issues. Moreover, the machine performance is commensurate with the human in terms of lexical diversity and coherence, as indicated by close scores of type-token ratio, trigram repetition, and sentence coherence. The results provide quantitative support that our model can generate grammatically correct and lexically coherent and diverse story texts.

**3.2.2 Human Evaluation.** We invite eight PhD students (four females and four males, mean age: 25.50 (SD = 2.07)) with English paper publications to rate their perceived quality of these 40 stories in random order. According to previous work [23, 89], we consider:

<sup>4</sup>Cosine similarities (range 0-1) between sentence embeddings using *sbert*.

**coherence** (*The story is logically consistent and coherent*), **relevance** (*The story is relevant to the title*), **interestingness** (*The story is interesting*), and **overall quality** (*Overall, it is a good story*). Each aspect is rated on a standard five-point Likert Scale (1 for “Strongly disagree” and 5 for “Strongly agree”). As shown in Table 3, the machine-generated stories achieve comparable performance with the human version regarding overall quality and interestingness. The human-written stories are considered significantly more coherent and relevant using the Wilcoxon Signed-rank test. Nevertheless, the machine-generated stories have average scores of around four points in terms of coherence and relevance. Therefore, we consider that our system could produce adequate story context given a set of target words for vocabulary learning.

### 3.3 Exploring Vocabulary Learning Activities with Story Generation Models

After validating the feasibility of our model for generating meaningful context that covers a set of CET-4 English words, we explore possible vocabulary learning activities that the model can support. We conduct semi-structured interviews with three English teachers (E1-3, age: 27 - 28) and five university students (S1-5, age: 21 - 29) in China. E1 has two years of experience in teaching IELTS and half-a-year experience in teaching English in a higher vocational college. E2 has spent five years in high-school English teaching, and E3 has taught high-school students mainly about TOEFL writings for three years in an educational institution. S1-5 are well-experienced in using different English vocabulary learning software for Chinese (*e.g.*, Liulishuo, Baicizhan, Shanbay).

**3.3.1 Procedure.** Each interview starts with participants’ practices (whether, why, and how) of story-based activities for teaching or learning vocabulary. Then, we show participants a web interface that allows users to input target English words and generate stories

with those words based on our model. We prepare example CET-4 word sets, each with the top-five topic-relevant words (e.g., cable, complain, library, instruction, unfortunate)<sup>5</sup> under our specified titles (e.g., the internet) and the generated stories in the interface. We invite our participants to check the generated stories and have a trial using their specified words. During this process, we encourage them to brainstorm the vocabulary learning activities that our story generation model can support. Each interview lasts for about 30 minutes with about USD \$3.5 for compensation.

**3.3.2 Results.** We transcribe the audio data into texts and group them into themes following the interview structure. Both groups of interviewees confirmed that learning English vocabulary via stories is a common and effective practice. For example, E1 mentioned that he usually asks students to first write sentences and then create a short story with newly learned words, which helps them master the usage of words. Three student participants regularly read English books and articles, which expands their vocabulary. In general, all participants agreed that our generative stories are suitable materials for learning target words. For instance, E3 tried the story generation model using “health” as title and “tobacco, alcohol, abuse, dominate, harmful” as target words. These words come from an article of her high-school text book. *“I like the generated story. It is generally coherent, and it is simpler than the one in the text book. My students would like it for vocabulary learning as they do not need to pay too much attentions on the long sentences”* (E3). Nevertheless, our three teachers pointed out that the generated stories lack explicit logic transition words like “nevertheless” and “for example”, which can further improve the stories’ coherence. This is probably due to the lack of these words in our training dataset ROCStory corpus.

Our interviewees actively provide ideas for leveraging our generative model to support vocabulary learning. Together with the insights from pedagogical literature (e.g., those in subsection 2.1), we summarize three supported vocabulary learning activities.

**Story reading:** learners can read the generated story to understand the meanings of the target words (E1-3, S1-5). This is also a typical meaning-focused input activity suggested by language educators [47].

**Cloze test:** learners can do a cloze test that fills blanks of the generated story using target words to strengthen their understandings (E2, S2, S3). *“Cloze test is a common vocabulary learning strategy in the textbook. I feel that it would be helpful to customize the generated stories into cloze tests for students”* (E2). Cloze test can be viewed as a language-focused learning activity with a focus on the usage of target words [47].

**Turn-taking writing:** learners can take turns with the generative models to co-write a story using target words (E1, E3). This practice combines the meaning-focused output and fluency development learning activities, as it requires learners to use the learned words productively and fluently [47]. *“It can generate a sentence using a target word as a start, and users write down the second sentence using another word. With such interaction, students can learn how to use the words in a successive manner”* (E1). *“The system can act as one student to do a turn-by-turn, co-writing practice”* (E3). The system can provide in-situ guidance and feedback on users’ input

<sup>5</sup>We use sentence-bert [63] to encode the words into vectors and rank them based on their cosine similarities with the vector of the encoded title.

in the writing process [47], e.g., *“are target words used correctly”* (E3) and *“is the story coherent and correct in grammar?”* (E1).

## 4 PHASE 2: STORYFIER SYSTEM IMPLEMENTATION AND REFINEMENT

After validating the feasibility of our story generation model and identifying promising ways to apply it, we present our second-phase design process about how we implement *Storyfier* and refine it with feedback from learners, teachers, and HCI researchers.

### 4.1 First *Storyfier* Prototype with Three Modes

Based on the interview findings, we design and implement three modes of user interfaces to facilitate vocabulary learning via story reading (Figure 3B), cloze test (C), and turn-taking writing (D).

**4.1.1 Interface Designs and User Workflow.** All three modes share the following two features (Figure 3A). [**Target words setting**] Users can manually add new words (“+”) or delete them (“x”) as they wish. They can also click the **C** button to get a randomly sampled target word set. [**Dictionary lookup**] Users can click each target word to inspect its definition, part of speech, phonetic symbol, and usage example. The click on  will lead to three vocabulary learning activities supported in the following interface variants.

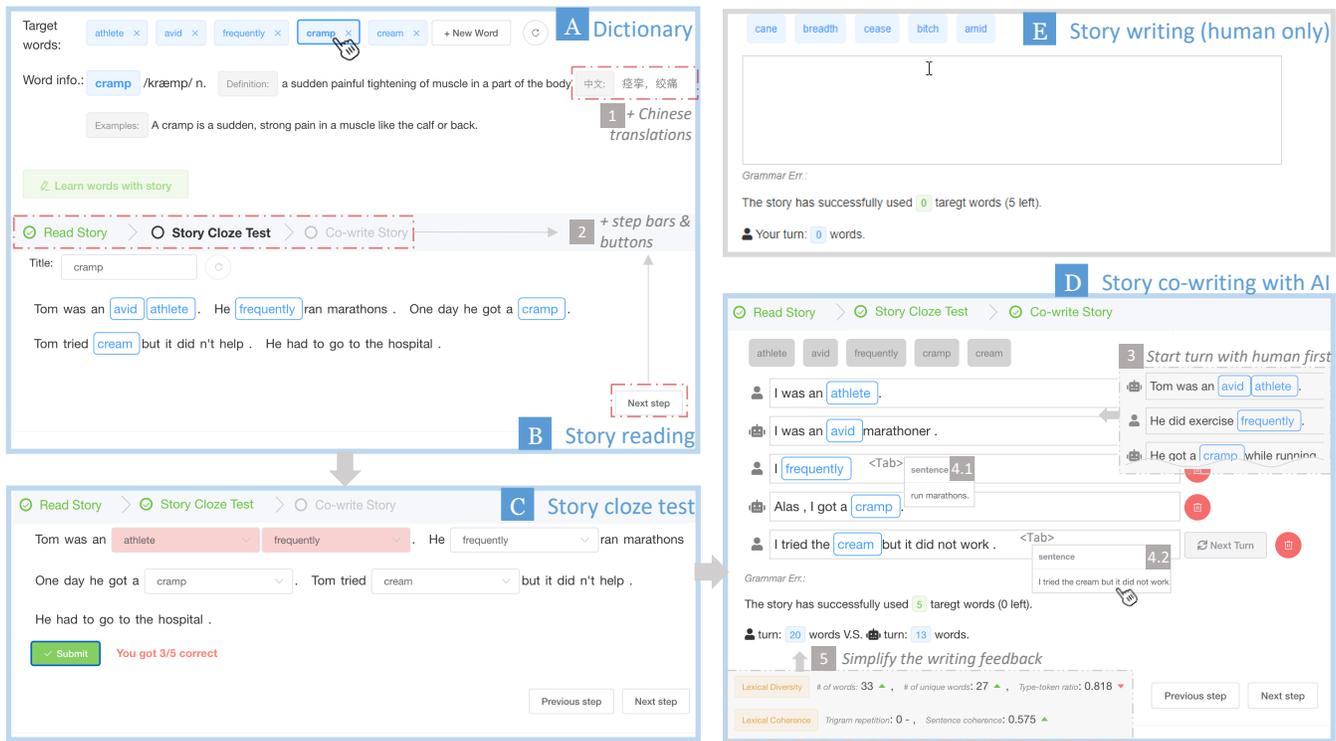
**Story reading mode.** This interface (Figure 3B) presents the AI-generated story with the target words highlighted in blue, which could help users quickly inspect their contextual use. In addition, users can conduct minor edits (e.g., , revise words) of the sentences to refine the story if they wish.

**Cloze test mode.** This interface (Figure 3C) replaces the target words in the generated stories with blanks. Users are required to make contextual inferences about the missing words and choose the proper ones to fill the blanks. After they submit the results, *Storyfier* will check the correctness and highlight the misused words (if any) in red. Users can iteratively fix the errors if they wish.

**Turn-taking writing mode.** This interface (Figure 3D) encourages users to write a story with AI using the target words sentence by sentence. During the writing process, users can gain an overview of the used (gray) and unused (blue) target words at the top in Figure 3D. The used target words are highlighted in the corresponding sentences. To provide adaptive feedback to learners, in each turn, the system will check and alert the grammar issues of the written text using LanguageTool API<sup>6</sup>. Meanwhile, *Storyfier* provides writing feedback on the story sentences at the bottom regarding grammar errors, lexical diversity, and lexical coherence (in Figure 3-5). Red and green triangles indicate a decrease or increase in scores of all current story sentences compared to the one in the previous turn. Users can write and refine the story with *Storyfier* until all the target words are used.

**4.1.2 Controllable story generation model that supports turn-taking writing.** To support the turn-taking writing activity where *Storyfier* needs to produce varied-length text spans given previous story sentences and target words, we further build a story-infilling model (Figure 2B-2). We formulate the training objective as a span prediction task and adopt a prompt-based approach to finetune the T5 model. Given a ROC story, we derive its story title, target words,

<sup>6</sup><https://languagetool.org/http-api/>



**Figure 3: The interface designs of *Storyfier*. (A) Users can specify target words and check their meanings. (B) Story reading: users can read a machined-generated story that contains target words. (C) Cloze test: users can conduct a cloze test by using target words on the generated story. (D) Story writing: users can take turns with *Storyfier* to write a new story using target words. (E) The interface for story writing without adaptive support in the *Storyfier*-sen baseline (section 5). Note that in the first *Storyfier* prototype, the three modes (B-D) are separated. In the refined *Storyfier*, we unify them into one flow and improve the system designs (1-5) based on feedback from learners and experts.**

and story sentences as prompts (described in subsection 3.1.2). Meanwhile, for each story sentence, we randomly mask varied-length of text spans of this sentence, following prior work [15]. Then, we train the model to predict the masked spans of the current sentence based on the prompts. We use a cross-entropy loss and finetune the pretrained T5-large model provided by Huggingface on our mask prediction task using adam optimization algorithm. We train 10 epochs, and the training loss is 1.0701.

With this finetuned model, *Storyfier* can write the next sentence using target words following the users’ written ones. Furthermore, in our refined *Storyfier* presented below (subsection 4.3), it can also help users revise an existing sentence or complete the unfinished one via text infilling.

## 4.2 Testing *Storyfier* Prototype

To seek feedback on the 1st *Storyfier* prototype, we conduct a usability test with ESL learners and two workshops with English teachers and HCI researchers.

**4.2.1 Usability Test with 12 ESL Learners.** To probe the user experience and perceived usefulness of the three activities supported by *Storyfier*, we conduct a within-subjects usability test with 12 junior undergraduate students (6 females, 6 males, mean age: 19.5 (SD =

0.52)) in a university in China. The baseline condition does not have the generated story contexts but provides a *dictionary* function that shows the meanings, synonyms/antonyms, and usage examples for each target word (Figure 3A1). All participants have passed the national English exam CET-4, with an average score 560.50 (SD = 37.75)<sup>7</sup>. We do not aim to evaluate *Storyfier*’s effectiveness but seek to improve it with quick user feedback at this stage. These participants can provide us with valuable feedback as they have fresh CET-4 vocabulary learning experience.

[*Procedure*] Participants use their own computers to remotely conduct the study following the instructions. They experience the four experiment conditions (i.e., *Dictionary*, *Read*, *Turn-taking Write*) one by one in a Latin-Scale counterbalanced order. In each condition, they learn two prepared word sets, each with five CET-4 words sampled based on topic relevance. After each condition, participants rate their perceived usefulness, easiness to use, and intention to use [79, 81] of each interface in a 7-points Likert scale; 7 for a strong agreement. In the end, we ask for their comments and suggestions on *Storyfier*. They receive about USD \$9.5 for around 50 minutes spent in the study.

<sup>7</sup>425/710 points are considered passed for CET-4.

**[Results]** We use repeated measured ANOVA test (*Dictionary vs. Read vs. Cloze vs. Turn-taking Write*) to evaluate the user experience of *Storyfier*'s three modes. There is a significant difference in perceived usefulness of the four activities;  $F(3, 33) = 3.83, p < 0.05, \eta^2 = 0.26$ . Specifically, they feel that the *Read* ( $M = 4.94, SD = 1.15, p < 0.05$ ) system is significantly more useful than the *Dictionary* one ( $M = 3.56, SD = 1.45$ ). Participants feel that the *Read* ( $M = 4.83, SD = 1.12, p < 0.01$ ) system is significantly easier to use than the *Turn-taking* one ( $M = 3.15, SD = 0.85$ );  $F(3, 33) = 9.54, p < 0.001, \eta^2 = 0.46$ ; Bonferroni post-hoc test. Besides, the *Cloze* ( $M = 4.27, SD = 1.13, p < 0.05$ ) system is deemed significantly easier to use than the *Turn-taking* one. Lastly, participants have significantly higher intentions to use the *Read* ( $M = 4.71, SD = 0.33$ ) system for their vocabulary learning in the future, compared to the *Dictionary* ( $M = 2.96, SD = 1.63, p < 0.01$ ) and *Turn-taking* systems ( $M = 3.13, SD = 1.40, p < 0.05$ );  $F(3, 33) = 5.80, p < 0.01, \eta^2 = 0.35$ . In summary, ESL learners found that the three learning activities supported by *Storyfier* are more useful than the baseline without story context. However, the *Storyfier*'s turn-taking writing mode should be further improved. For example, two students indicated that sometimes they found it difficult to use words to write the next story sentence in this activity.

**4.2.2 Co-Design Workshops with English Teachers and HCI Researchers.** Apart from the feedback from ESL learners, we conduct two co-design workshops to seek experts' feedback. The two workshops share a similar procedure but have a different focus. One is with the same three English teachers (E1-3) in Phase 1 and focuses on refining the vocabulary learning activities in *Storyfier*. The other is with four HCI researchers (H1-4, all males, age: 25-27) and mainly works on the interface and interaction design of *Storyfier*. All HCI researchers have experience in developing intelligent systems and have papers published in top venues like CHI and VIS. Each workshop starts with a warm-up activity in which participants share their experience of story-based vocabulary teaching or learning. Then, we show our *Storyfier* prototype to them, invite them to have a trial, and ask them to give comments on the system. Next, we organize a brainstorming session to discuss how to leverage the three learning activities of *Storyfier* for effective vocabulary learning support and how to improve the interaction and interface design. Each workshop lasts about one hour, and participants receive about USD \$17 as compensation. We present their suggestions on *Storyfier* together with its refinement in the next subsection.

### 4.3 Refined *Storyfier* System

Based on the collected feedback on the first *Storyfier* prototype, we refine its workflow and features (Figure 3).

**Workflow.** We unify the three separate learning activities into one workflow using step bars and next-step buttons (in Figure 3-2) to guide learners to read the story, do a cloze test, and write a new story. Our English teachers agree that all three learning activities would be generally helpful, but there could be a flow that chains these activities to maximize their values. They suggest that *reading* should be the first activity to help comprehend the target words. The cloze test should come next to strengthen their understanding, and the co-writing practice should be the last activity. "*Cloze test is a controlled practice, and co-writing is a free one*" (E3). To chain the

three learning activities into a flow, S3 proposes to use a chatbot to guide users through the learning process, which could be engaging. This is similar to the chatbot interaction in StoryBuddy [92]. However, the other three HCI researchers are concerned that it might distract users' attention from vocabulary learning to interaction with the chatbot. S1 suggests that we can use clear widgets (e.g., the right arrow and "Next Turn" buttons) to order the flow of the three activities.

**Features.** First, we add the main Chinese meaning of each target word in the dictionary (Figure 3-1) as suggested by E2. Second, we modify the turn-taking order by encouraging users write the first sentence of the story (Figure 3-3), as suggested by E3 that we should encourage learners to spend effort first. Third, we add an inline sentence suggestion function that can infill a generated next sentence using target words (Figure 3-4.1 and -4.2), to address learners' difficulties in story writing activity found in the usability test. This function can be triggered in real-time by the "tab" key on users' demands, as suggested by the HCI researcher H4. Third, we remove the technical metrics about sentence quality (Figure 3-5), as suggested by E1-3 that they are complicated for learners and not focused on the target words' usage. Fourth, we add the number of used target words and the number of words written by human/machine as writing feedback because it could encourage learners to write more.

## 5 EXPERIMENT

To explore how would *Storyfier* impact the users' vocabulary learning outcome and experience, we conduct an experiment with 28 ESL (English-as-the-Second-Language) Chinese students. We adopt a 2 (with vs. without AI features)  $\times$  2 (read-only vs. read-cloze-write activities) within-subjects design. The first one – **AI factor** – aims to study the impacts of *Storyfier*'s AI-generated story and adaptive writing support. We note the conditions with AI features with "-AI" and those without AI features with "-sen". The second one – **activity factor** – identifies the value of additional cloze test and writing activities to the reading activity that previous vocabulary learning support systems focus on. We note the conditions with the read-only activity with "Read-" and those with read-cloze-write activities with "Storyfier". The four conditions are:

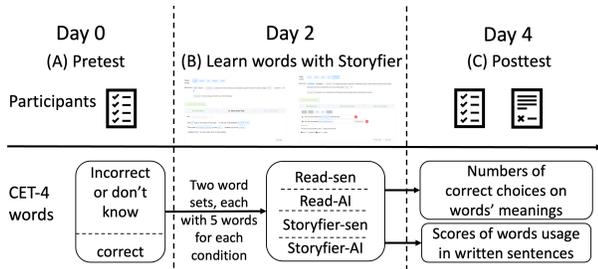
### Read-only

- **Read-sen** interface only provides dictionary features with an example existing sentence for each target word (Figure 3A);
- **Read-AI** interface additionally provides a generated story that covers target words (Figure 3A + B);

### Read-cloze-write

- **Storyfier-sen** interface offers example sentences for target words, a cloze test on these sentences, and a writing exercise without AI's intervention (Figure 3A + B + D, but the stories in B and C are replaced by the example sentences of target words);
- **Storyfier-AI** interface contains all features of *Storyfier* (Figure 3A + B + C).

The Read-sen and Storyfier-sen interfaces simulate how individuals traditionally use existing materials to learn any target word set



**Figure 4: Procedure of the experiment. (A) Participants first took a pretest, and the words they did not know were target words. (B) On the experiment day, they used the four interfaces of Storyfier for vocabulary learning. (C) Two days later, they took the posttest on words’ meanings and usage.**

without adaptive support, which can help us evaluate the impact of *Storyfier*’s story generation model.

Our research questions are:

**RQ1.** How would *Storyfier* affect vocabulary learning outcome?

**RQ2.** How would *Storyfier* affect the learning experience?

**RQ3.** What are user perceptions towards *Storyfier*?

## 5.1 Participants

We recruit 28 second-year undergraduate students (P1-28, 24 females, 3 males, 1 prefer not to tell, mean age: 20.04 ( $SD = 0.69$ )) from a course in a college in mainland China. They are typical ESL learners who major in Business English. The course nature leads to the gender and major unbalance of our participants, which we will discuss in the Limitations subsection. Twelve of them have not passed the national English exam CET-4 in China, and the rest have passed it with an average score 493.8/710 ( $SD = 35.8$ )<sup>8</sup>. Their self-assessed English vocabulary proficiency score is 4.39 ( $SD = 0.63$ ; 1 - not proficient at all, 7 - very proficient).

## 5.2 Procedure and Tasks

We conduct the experiment remotely. In a similar manner as [1], the procedure of our experiment consists of three stages (Figure 4). First, after collecting the background information with consent, we ask each participant to take a pretest to identify the CET-4 words that they did not know. In the pretest, the participant needs to choose one of the five options, including four meanings written in Chinese and one indicating “I do not know this word”, for each CET-4 word. We invite a postgraduate to prepare 170 CET-4 words that are not easy (e.g., excluding words like “easy” and “feel”) from the English learning app *Baicizhan* and only include the intended participants who answer incorrectly or indicate lack of knowledge on at least 40 words. For each participant, we randomly select 40 of the identified unknown words and divide them into eight sets, each with five target words. After the pretest, we also have participants read the instructions of the learning tasks and the four interfaces. We inform them not to learn the words that appear in the pretest prior to learning sessions.

<sup>8</sup>425/710 points are considered passed for CET-4.

Then, on the experiment day, participants log in to their learning sessions via their unique IDs. Participants are asked to learn two word sets with each *Storyfier* interface. We counterbalance the order of the four interfaces using Latine Square. After learning two word sets with an interface, participants rate their engagement and enjoyment in the learning process, perceived learning performance, and perceptions of the system in a questionnaire. Upon completion of four tasks, we further ask for their preferences on the interfaces, comments on the generated stories and AI’s writing assistance, and suggestions for improving *Storyfier*.

Next, two days after the experiment day, we ask the participants to take a posttest, which has a similar format as the pretest but only presents the 40 words they met in the learning sessions. In the posttest, participants also need to write a sentence for each target word if they do not choose “I do not know this word”. They can write “nothing” if they feel hard to write the sentence. Each participant spends about 1.5 hours in total on the full procedure and gets around USD \$12 as compensation.

## 5.3 Measurements

**RQ1. Learning Outcome.** We measure participants’ retention of target words’ meanings via the number of correct answers to the multiple-choice questions in the posttest. To capture how well they learn the usage of target words, we invite one English teacher (E1 in our workshop) to rate the grammar correctness (e.g., tense and part of speech) and context appropriateness of the target word in each written sentence in the posttest using a three-point scale; 0 - not correct, 1 - partially correct, 2 - correct. For each participant in each system interface, we calculate i) the numbers (range: 0 - 10) of sentences that use target words correctly in terms of both grammar and context and ii) the total score (0 - 40) of sentences<sup>9</sup>.

**RQ2. Experience.** We measure users’ engagement and enjoyment during the learning process with each system interface (“*I was absorbed in using this interface to learn vocabulary*” and “*It is enjoyable to learn vocabulary with this interface*” [80, 85]). Besides, we measure the perceived task workload of learning sessions using items adapted from NASA Task Load Index [28] (e.g., “*I have to work hard to accomplish the writing activity.*”). Apart from the questionnaire data, we also log the i) task completion time of learning two word sets with each interface, as well as ii) the amount of time spent in reading, cloze-test, and writing activities and iii) written stories in *Storyfier-sen* and *Storyfier-AI* interfaces.

**RQ3. Perceptions towards *Storyfier*.** We adapt the technology acceptance model [79, 81] to the perceived usefulness (four items, e.g., “*The use of this interface enables me to learn the vocabulary more efficiently*”; Cronbach’s  $\alpha = 0.944$ ), easiness to use (four items, e.g., “*I would find this interface to be flexible to use*”;  $\alpha = 0.786$ ), and intention to use (two items, e.g., “*If this interface is available there to help me learn vocabulary, I would use it*”;  $\alpha = 0.966$ ) of each system. We average the ratings of multiple questions as the final score for each aspect. All statements in the questionnaire are rated on a standard 7-point Likert Scale, with 7 for a strong agreement.

<sup>9</sup>In each interface, participants learn ten words and write at most ten sentences in posttest. The maximum score for each sentence is  $2 + 2 = 4$ .

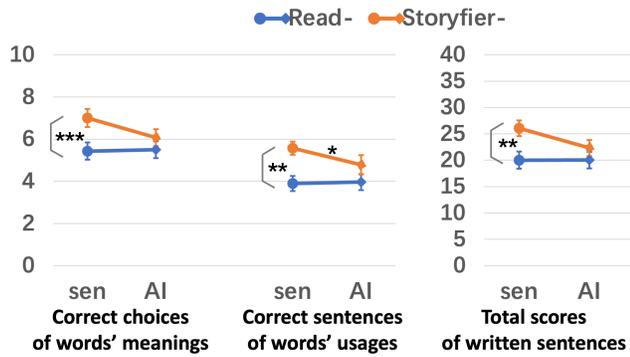


Figure 5: RQ1 results regarding numbers of correct choices on target words' meanings, numbers of sentences that correctly use target words, and total scores of the written sentences in each condition. \*\*\*:  $p < 0.001$ , \*\*:  $p < 0.01$ , \*:  $p < 0.05$ .

## 6 ANALYSES AND RESULTS

For the rated items, we first check whether the order of the four experienced interfaces affects our results via a set of mixed ANOVA tests (order as between-subjects, interfaces as within-subjects) on each rating. Neither the main effect of the order nor its interaction effect with the system interface is significant. Hence, except those with additional notations (e.g., one-way ANOVA), the statistic tests in this section are two-way repeated measured ANOVAs. For each ANOVA, the assumption of equal variance holds according to Macuchly's test of sphericity [21]. For the participants' comments on *Storyfier*, two authors conduct an inductive thematic analysis [3]. They first independently assign codes to the text data and then discuss the codes for several rounds. After that, they group the codes into categories, which are incorporated into the results below.

### 6.1 RQ1: Impact on Learning Outcome

Figure 5 shows the results regarding learning outcomes.

**6.1.1 Retention of target words' meanings.** Our results indicate that neither the AI factor nor its interaction with the activity factor significantly affects the retention of target words in four conditions. However, the *Storyfier*-sen interface results in a better retention performance ( $M = 7.00$ ,  $SD = 2.16$ ) than the *Storyfier*-AI interface ( $M = 6.07$ ,  $SD = 2.09$ );  $p = 0.049$ , one-way repeated-measures ANOVA. Besides, participants perform significantly better in target words' retention in the read-cloze-test (i.e., *Storyfier*-sen and *Storyfier*-AI) conditions ( $M = 6.54$ ,  $SD = 2.19$ ) than that in the read-only (i.e., *Read*-sen and *Read*-AI) conditions ( $M = 5.46$ ,  $SD = 2.16$ );  $F = 9.605$ ,  $p = 0.004$ .

**6.1.2 Target words' usage in the sentences.** Neither the AI factor nor its interaction with the activity factor has a significant impact on i) the number of sentences that correctly use target words and ii) the total scores of written sentences in four conditions. However, when comparing the means between the *Storyfier*-sen and *Storyfier*-AI interfaces, we observe that in the read-cloze-write learning sessions, *Storyfier*'s AI features could reduce learning gains on target words' usage. As for the activity factor, our results show that participants

with the read-cloze-write interfaces ( $M = 24.20$ ,  $SD = 8.66$ ) perform significantly better in word usage in their written sentences than the cases with the read-only interfaces ( $M = 20.02$ ,  $SD = 7.92$ ); e.g., for ii) total scores,  $F = 12.721$ ,  $p = 0.001$ .

In all, we find that *Storyfier*'s AI features reduce learning gains on the retention of target words' meanings in the read-cloze-write vocabulary learning sessions. Its supported additional cloze-test and writing practices improve learning gains on target words' meanings and usage compared to learning via reading-only activities.

### 6.2 RQ2: Impact on Learning Experience

**6.2.1 Engagement, enjoyment, and workload.** As shown in Figure 6<sup>10</sup>, neither the AI factor nor its interaction with activity factor significantly affects users' perceptions on their learning experience. When digging into each measured item in each condition, we have several interesting observations: a) in read-only sessions, participants with AI-generated stories could feel more engaged and enjoyed than the cases without these stories; b) *Storyfier*'s AI features could increase mental demand and perceived performance in read-only learning sessions but decrease the ratings on these measures in read-cloze-write sessions; c) *Storyfier*'s AI features could reduce temporal demand, i.e., how rushed is the pace of the task, and perceived spent effort in the vocabulary learning tasks. As for the activity factor, we found significant differences regarding the perceived mental demand ( $p = 0.002$ ), physical demand ( $p = 0.029$ ), and perceived performance ( $p = 0.025$ ). Specifically, *Storyfier*'s supported additional cloze-test and writing practices increase mental and physical demand and perceived performance compared to learning via reading-only activities. We also observe that these additional activities could increase engagement and spent effort in the vocabulary learning tasks.

**6.2.2 Task completion time and written stories.** i) On average, participants spent 89.22( $SD = 105.25$ ) / 226.52(150.66) / 805.00(490.55) / 806.61(368.19) seconds in the learning session with *Read*-sen, *Read*-AI, *Storyfier*-sen, or *Storyfier*-AI interface. This indicates that in the read-only vocabulary learning sessions, participants spent significantly more time when they were presented with AI-generated stories than when they were not ( $p < 0.001$ ). However, as shown in Figure 5, the learning gains do not increase accordingly. ii) When digging into the average amount of time spent in each learning activity for each word set in *Storyfier*-sen and -AI interfaces, we have 49.48 vs. 81.56 (read,  $p = 0.004$ ), 30.76 vs. 43.84 (cloze,  $p = 0.004$ ), and 263.46 vs. 208.71 (write,  $p = 0.09$ ) seconds<sup>11</sup>. This shows that compared to the sessions with *Storyfier*-sen, participants with *Storyfier*-AI spent significantly more time in reading and cloze-test activities but less time in writing activities.

### 6.3 RQ3: Perceptions on *Storyfier*

Figure 7 depicts users perceptions of each *Storyfier* interface.

**6.3.1 Quantitative items.** In read-only sessions, there is a trend on the improved perceived usefulness and intention to use of the interfaces with AI-generative stories over those without the stories.

<sup>10</sup>The full statistics are attached in the supplementary materials.

<sup>11</sup>The total time does not match task completion time as it does not include time spent on checking each word's meaning.

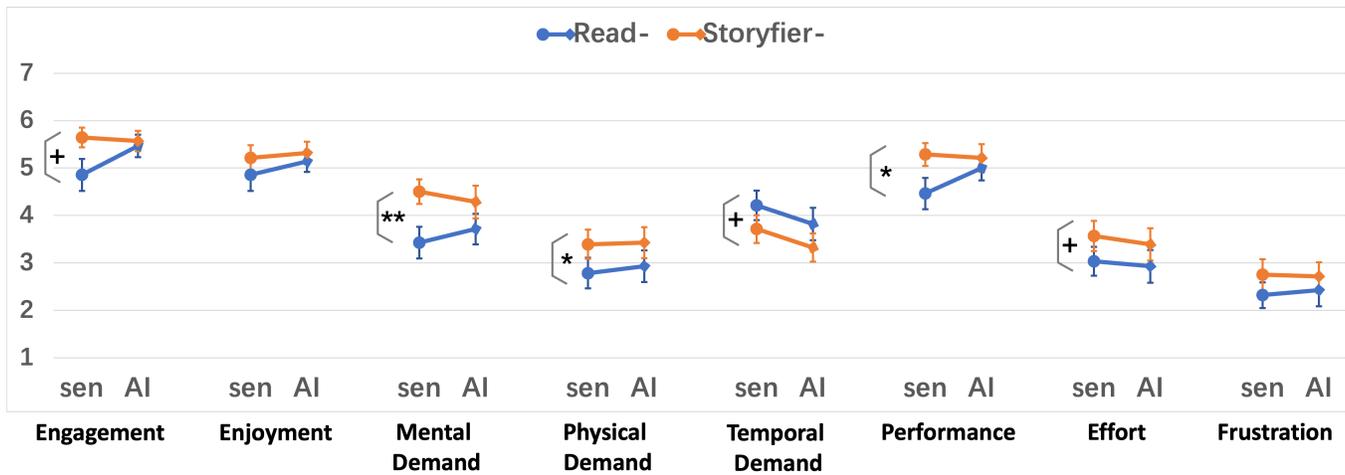


Figure 6: RQ2 results regarding perceived engagement, enjoyment, and workload in vocabulary learning sessions with Read-sen, Read-AI, Storyfier-sen, and Storyfier-AI interfaces. \*\*:  $p < 0.01$ , \*:  $p < 0.05$ , +:  $p < 0.1$ .

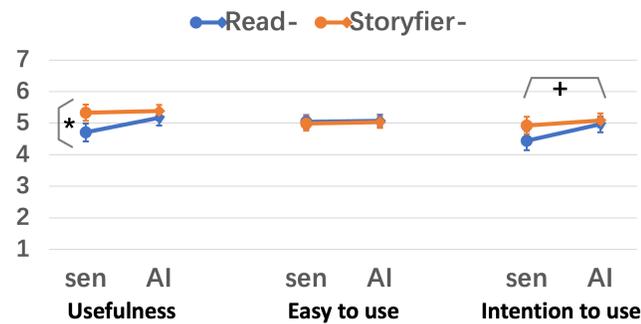


Figure 7: RQ3 results regarding user perceptions with each interface. \*:  $p < 0.05$ , +:  $p < 0.1$ .

This implies that in read-only learning sessions, participants would find Storyfier more useful and have a higher intention to use it if it provides AI-generated stories. As for the activity factor, participants feel that *Storyfier* is significantly more useful ( $p = 0.035$ ) if it supports cloze test and writing practices in addition to the reading activities. There are no significant differences in the perceived easiness to use across the four interfaces.

**6.3.2 Qualitative responses. Preference.** In the open-response questions after four learning sessions, fifteen participants indicate their preferences for the Story-AI interface for vocabulary learning. They especially favor adaptive writing assistance ( $N = 9$ ), generative stories (5), and useful practices (4). “It (Storyfier-AI) not only provides the meaning, pronunciation, and example sentences of words, but more importantly, it has AI-generated short stories that can help me better understand the meanings of words and how to use them. In addition, the following cloze test and story writing practice can further consolidate my understanding. When I do not know how to write, AI will also provide prompts to help me find my weak points and mistakes so that I can pay more attention on them later on” (P22).

Six participants prefer the Story-sen interface, and three of them credit the writing practice without AI assistance. “I prefer Storyfier-sen as I need to rely on myself to think and write down the story, which would be more impressive for vocabulary learning” (P8). Five participants prefer the Read-AI interface for its low task workload ( $N = 3$ ), meaningful contexts for learning ( $N = 3$ ), and enjoying the experience (2), while the rest two participants favor the Read-sen one as they are more used to the rote learning practices.

**Generative stories.** Regarding participants’ comments on the generative stories, we found positive opinions that they are coherent (8) and novel/interesting/impressive (9). P26 gives us an example. “At first, I could not remember the word ‘veil’. Then, I checked the generated story, which tells that a veil blocked my vision when I was driving in traffic. This story is close to real life, and I felt terrified. It is impressive. I remembered the word ‘veil’ now.” Nevertheless, there are six comments suggesting that the stories were not coherent, which may be due to the lack of semantic connections among the target words. “When the five target words, e.g., ‘hasten, infinity, jet, basin, and trolley’ are not naturally relevant to each other, it would be hard to have a reasonable story that covers them, making it hard for memorizing the words in a batch” (P27). Besides, three users comment that some target words in the generated stories have different meanings from the dictionary ones, and another three users mention that the stories contain some words that are unknown, which disturbs story comprehension.

**Cloze test.** Twenty users indicate their preferences on using generative stories for the cloze test, which can “make it easy to connect the words in context” ( $N = 10$ ), “enhance memory of target words” (7), and “train reading comprehension skills” (2). The other eight participants, however, prefer the existing sentences provided by Storyfier-sen for cloze test materials, with four comments mentioning that “separate sentences are easier to understand”.

**Writing practice.** There are seventeen positive responses on the adaptive writing assistance from the generative AI models, suggesting that it can encourage writing (8), reduce writing workload

(5), and provide example usage of target words for reference (4). “I didn’t feel confused when writing with Storyfier-AI. I can write a sentence first and then let the AI write the next one, and so on. It’s like having a buddy to memorize words together, which is more interesting and not boring” (P2). “The AI’s prompts inspire my writing exercises” (P23). However, these prompts in the turn-taking writing process may not match the learners’ idea flow (2) and language styles (1) and cause their reliance on AI for using target words (1).

In all, these qualitative responses reveal that participants generally favor Storyfier with AI generative models in Read-AI and Storyfier-AI learning sessions compared to the Read-sen and Storyfier-sen interfaces. However, these models still need to be improved and customized regarding the coherence, complexity, and style of the generative content.

## 7 DISCUSSION

### 7.1 Insights from our findings

**7.1.1 Text generation models for vocabulary learning support.** In Phase 1, we develop a story generation model and verify that it can generate comparably good stories with the human-written ones given target word sets and titles. We receive divided opinions on the generative stories regarding their coherence and interestingness in the experiment with 28 ESL learners. The main reason could be that if the target five words are not naturally relevant to each other, it would be difficult for the generative model to connect them to form a meaningful story. Besides, we get feedback from English teachers in Phase 1 that our generative stories have generally acceptable complexity for vocabulary learners. Yet, there are still cases that our stories contain unknown words in addition to the target ones.

**7.1.2 Vocabulary learning activities.** In Phase 1, we propose that our generative models can empower the cloze test and writing practices in addition to the traditional reading activities that previous vocabulary learning tools support. Our evaluation study with ESL learners reveals that they have significantly more learning gains in read-cloze-write (*i.e.*, Storyfier-sen and Storyfier-AI) learning sessions compared to that in read-only (*i.e.*, Read-sen and Read-AI) sessions. This is non-surprising as learners spent more effort in understanding target words and practicing their usage in the read-cloze-write sessions. Participants’ responses in the open-ended questions suggest that generative models can facilitate vocabulary learners by providing meaningful reading and cloze-test materials and adaptive prompts in writing practices.

**7.1.3 Impact of generative models on vocabulary learning.** In the read-only sessions, we observe that the additional AI-generated stories improve learning engagement but do not improve learning gains. This does not support the Gu *et al.*’s implication that learning vocabulary in a batch under a coherent context could facilitate recalls of target words [25]. One key reason could be the low semantic relevance among some target words in our experiment. Gu *et al.*’s implication could still be valid if the selected target words are topically relevant to each other, *e.g.*, as organized in a typical language course book. Generative stories can explicitly reveal the words’ relationship.

In the read-cloze-write sessions, we found that the AI support leads to reduced vocabulary learning gains. We attribute these

results to the amount of effort spent in the writing practices. While participants mostly favor the assistance from our generative model, they spent less time in writing and wrote significantly fewer words in the story. This provides a lesson that the AI’s assistance should encourage necessary effort in vocabulary learning instead of aiming to reduce learners’ workload.

### 7.2 Design Considerations

Based on our findings, we outline three directions for supporting vocabulary learning with generative models.

**Support four strands of vocabulary learning activities with generative models.** Our results (subsection 6.1) with 28 ESL learners show that *Storyfier* improves learning gains compared to the read-only baselines. This improvement can be due to the *Storyfier*’s cloze test and writing activities that integrate the recommended four strands of learning activities [47] (subsection 2.1). We thus recommend that vocabulary learning tools should integrate multiple strands of activities. While prior language learning systems, *e.g.*, Smart Titles [36] and EnglishBot [66], have explored vocabulary learning activities like watching videos and speaking to others, they largely leveraged existing learning materials. We suggest that generative techniques, such as the story generation model we developed, the text-to-image [87], the text-to-video [42], and the music generation [29] approaches, can enrich the learning materials and offer in-situ assistance in these vocabulary learning activities. For example, the learning support tool can generate an image based on the example sentence of each target word to help them understand the word’s meaning. It can further offer generated music clips that use this sentence as listening resources.

**Provide adaptive learning feedback.** *Storyfier* offers feedback on the correctness of cloze test results and grammar of written sentences. However, two participants comment that it could offer more adaptive learning feedback. For example, it can “*first let the learner draft a story and then assess its quality and usage of target words*” (P27). It can further “*recommend how to improve the written story*” (P25). Previous skills learning tools, such as ArgueTutor [80], Persua [86] and VoiceCoach [83], and writing support tools like MepsBot [56] also adopt a similar feedback flow, which can mitigate disturbance on the practicing process. Generative models can offer such learning feedback by generating polished versions of the written story for reference. However, based on the teachers’ suggestions on our first *Storyfier* prototype (Figure 3D-5), the provided feedback should emphasize the vocabulary learning goal; otherwise, learners may chase for other objectives, *e.g.*, trigram repetition and sentence coherence.

**Balance machine and human effort in learning tasks.** Our results show that *Storyfier*’s AI features reduce learning gains on the retention of target words’ meanings in the read-cloze-write sessions (subsection 6.1). Participants report the need for increased workload and autonomy in the writing task for effective vocabulary learning (subsubsection 6.3.2). As such, we recommend that *Storyfier* should further motivate necessary user effort in learning. In the refinement of *Storyfier* (subsection 4.3), we have experienced a few features to encourage more user effort, *e.g.*, users need to write the first sentence of the story. Future work could explore the inclusion of

gamification features like badges, timers, and leader boards for promoting learners' efforts in educational scenarios [12].

### 7.3 Generality of *Storyfier*

While *Storyfier* presets the target words that participants are unknown in the experiment, we have equipped it with features like adding or deleting any words, suggesting words semantically related to the title, filtering words based on difficulty level, and editing the generated stories. In other words, *Storyfier* can support customized individual vocabulary learning beyond the controlled lab sessions. Besides, our English teachers in the design workshop express their interest in applying *Storyfier* in language teaching. One teacher, E2, had a trial on her offline course by inputting five words she just taught and asking students to have a cloze test on the generated story. Therefore, our *Storyfier* is promising for supporting customized vocabulary learning and teaching in the wild.

### 7.4 Limitations and Future Work

**Handle language ambiguity.** Currently, our system does not consider language ambiguity when generating stories for target vocabulary. For example, one word can carry multiple meanings (*i.e.*, polysemy) in different contexts, while our system only considers its most common meaning in practical usage. Comparative studies of the same word in different contexts can help disambiguate words and deepen the understanding of vocabulary.

**Improve story generation quality.** To further improve the quality of story generation for vocabulary learning, we can consider two aspects: dataset and model. The simplicity of the ROCStory dataset, while appreciated by English teachers for vocabulary learning, has certain limitations. It lacks transition words, which makes it challenging for models to learn sentence transition logic. Additionally, the simplicity of the story structures can potentially compromise the richness of intra-sentence contexts. In the future, we can incorporate more complex stories with a wide range of narrative structures into our dataset for model training.

Besides, we can investigate the use of more powerful language models (*e.g.*, ChatGPT) to enhance the quality of the generated stories. For example, we have experimented with prompts (*e.g.*, with “simple”, “CET-4”, “within 50 words”, and/or “no more complex words”) to steer ChatGPT towards generating stories that fulfill our specifications<sup>12</sup>. Notably, however, there is a trade-off between the simplicity and coherence of the generated stories. Future research can focus on refining prompting strategies to optimize the balance between these two elements for enhancing the efficacy of vocabulary learning.

**Generate diverse and harmless stories.** *Storyfier* presents one story at a time based on the generation models trained on ROCStory dataset, which contains simple and short stories. In the future, we can consider generating more diverse stories (*e.g.*, in the form of newspapers, novels, poems, and humor) that have varied styles and lengths for vocabulary learning. Besides, while our participants did not report harmful content in the generated stories, *Storyfier*

should include features like “report” and automatic detection of unwanted content to offer a healthy learning environment.

**Identify helpful characteristics of stories for vocabulary learning.** We evaluate the quality of our generated stories via coherence, relevance, and interestingness (Table 3) and collect qualitative feedback from participants on the stories' helpfulness. We call for future work to complement our evaluation studies by identifying what are the helpful characteristics of stories for vocabulary learning. For example, we can talk to English-learning textbook authors or conduct a content analysis on textbook stories. The identified characteristics can further guide us to customize story generation models and enhance our evaluation metrics of the stories.

**Invite diverse language learners.** We conduct the experiment with Chinese students in an English learning course to evaluate *Storyfier*. Their CET-4 test scores and self-reported proficiency indicate that they are intermediate-level English learners. Further studies can explore whether and how *Storyfier* with generative models can support novices with no or little prior experience in learning English. Moreover, we can extend *Storyfier* to support users from different cultures to learn their foreign languages (*e.g.*, English learners study Chinese).

**Evaluate cloze test and writing practices separately.** In the experiment, we evaluate the impact of the cloze test and writing practices by comparing the participants' performance and experience in read-only and read-cloze-write sessions. This study design is to identify the value of the vocabulary learning activities beyond the meaning-focused input activities that previous systems support. However, we can not quantitatively tell how much the cloze test or writing practice contributes to the impact, which requires a future study that separately evaluates these two activities.

## 8 CONCLUSION

In this paper, we designed and developed an interactive system, *Storyfier*, to support reading, cloze test, and writing activities for vocabulary learning. We power the system with controllable language models that can generate stories given any target words and provide adaptive assistance when using these words in the writing practices. We explore its supported vocabulary learning activities and interface design with teachers, learners, and Human-Computer Interaction researchers. Our two-by-two within-subjects experiment with 28 English-as-Second-Language Chinese students shows that participants generally favor the generated stories and writing assistance. However, their learning gains with *Storyfier* in the read-cloze-test sessions decrease compared to the cases they are with a baseline system without generative models. We discuss insights from our findings for leveraging generative models to support learning tasks.

## ACKNOWLEDGMENTS

This work is supported by the Young Scientists Fund of the National Natural Science Foundation of China with Grant No. 62202509 and partially supported by the Research Grants Council of the Hong Kong Special Administrative Region under General Research Fund (GRF) with Grant No. 16203421.

<sup>12</sup>For example, we queried ChatGPT using “write a five-sentences simple story using words: hasten, infinity, jet, basin, and trolley”. This results in a 71-word coherent story but contains more complex sentence structure and words like “marvel”, “exhilarated”, and “adventure”.

## REFERENCES

- [1] 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* (New Orleans, LA, USA) (CHI '22). Association for Computing Machinery, New York, NY, USA, Article 6, 21 pages. <https://doi.org/10.1145/3491102.3501839>
- [2] Gordon H Bower. 1970. Analysis of a mnemonic device: Modern psychology uncovers the powerful components of an ancient system for improving memory. *American Scientist* 58, 5 (1970), 496–510.
- [3] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. <https://doi.org/10.1191/1478088706qp0630a>
- [4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.
- [5] Daniel Buschek, Martin Zürn, and Malin Eiband. 2021. The impact of multiple parallel phrase suggestions on email input and composition behaviour of native and non-native english writers. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.
- [6] Alex Calderwood, Vivian Qiu, Katy Ilonka Gero, and Lydia B Chilton. 2020. How Novelists Use Generative Language Models: An Exploratory User Study.. In *HAI-GEN+ user2agent@IUI*.
- [7] Patricia L Carrell. 1984. Schema theory and ESL reading: Classroom implications and applications. *The modern language journal* 68, 4 (1984), 332–343.
- [8] Chih-Ming Chen and Ching-Ju Chung. 2008. Personalized mobile English vocabulary learning system based on item response theory and learning memory cycle. *Computers & Education* 51, 2 (2008), 624–645.
- [9] John Joon Young Chung, Wooseok Kim, Kang Min Yoo, Hwaran Lee, Eytan Adar, and Minsuk Chang. 2022. TaleBrush: Sketching Stories with Generative Pretrained Language Models. In *CHI Conference on Human Factors in Computing Systems*. 1–19.
- [10] Hai Dang, Karim Benharrak, Florian Lehmann, and Daniel Buschek. 2022. Beyond Text Generation: Supporting Writers with Continuous Automatic Text Summaries. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–13.
- [11] Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2019. Plug and play language models: A simple approach to controlled text generation. *arXiv preprint arXiv:1912.02164* (2019).
- [12] Paul Denny, Fiona McDonald, Ruth Empson, Philip Kelly, and Andrew Petersen. 2018. Empirical support for a causal relationship between gamification and learning outcomes. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–13.
- [13] Griffin Dietz, Jimmy K Le, Nadin Tamer, Jenny Han, Hyowon Gweon, Elizabeth L. Murman, and James A Landay. 2021. Storycoder: Teaching computational thinking concepts through storytelling in a voice-guided app for children. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.
- [14] Griffin Dietz, Nadin Tamer, Carina Ly, Jimmy K Le, and James A Landay. 2023. Visual StoryCoder: A Multimodal Programming Environment for Children's Creation of Stories. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–16.
- [15] Chris Donahue, Mina Lee, and Percy Liang. 2020. Enabling Language Models to Fill in the Blanks. *ArXiv abs/2005.05339* (2020).
- [16] Rita Stafford Dunn and Kenneth J Dunn. 1972. *Practical approaches to individualizing instructions: contracts and other effective teaching strategies*. Parker Publishing.
- [17] Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833* (2018).
- [18] Liye Fu, Benjamin Newman, Maurice Jakesch, and Sarah Kreps. 2023. Comparing Sentence-Level Suggestions to Message-Level Suggestions in AI-Mediated Communication. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 103, 13 pages. <https://doi.org/10.1145/3544548.3581351>
- [19] Pablo Gervás, Belén Díaz-Agudo, Federico Peinado, and Raquel Hervás. 2004. Story plot generation based on CBR. In *International Conference on Innovative Techniques and Applications of Artificial Intelligence*. Springer, 33–46.
- [20] Marjan Ghazvininejad, Xing Shi, Jay Priyadarshi, and Kevin Knight. 2017. Hafez: an interactive poetry generation system. In *Proceedings of ACL 2017, System Demonstrations*. 43–48.
- [21] E.R. Girden. 1992. *ANOVA: Repeated measures*. Sage Publications, Inc., Thousand Oaks, CA, US.
- [22] Seraphina Goldfarb-Tarrant, Tuhin Chakrabarty, Ralph Weischedel, and Nanyun Peng. 2020. Content planning for neural story generation with aristotelian rescoring. *arXiv preprint arXiv:2009.09870* (2020).
- [23] Seraphina Goldfarb-Tarrant, Haining Feng, and Nanyun Peng. 2019. Plan, Write, and Revise: an Interactive System for Open-Domain Story Generation. In *NAACL*.
- [24] Steven M Goodman, Erin Buehler, Patrick Clary, Andy Coenen, Aaron Donsbach, Tiffanie N Horne, Michal Lahav, Robert MacDonald, Rain Breaux Michaels, Ajit Narayanan, et al. 2022. LaMPost: Design and Evaluation of an AI-assisted Email Writing Prototype for Adults with Dyslexia. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility*. 1–18.
- [25] Yongqi Gu and Robert Keith Johnson. 1996. Vocabulary Learning Strategies and Language Learning Outcomes. *Language Learning* 46, 4 (1996), 643–679. <https://doi.org/10.1111/j.1467-1770.1996.tb01355.x>
- [26] Yongqi Gu and Robert Keith Johnson. 1996. Vocabulary learning strategies and language learning outcomes. *Language learning* 46, 4 (1996), 643–679.
- [27] Brent Harrison, Christopher Purdy, and Mark O Riedl. 2017. Toward automated story generation with markov chain monte carlo methods and deep neural networks. In *Thirteenth Artificial Intelligence and Interactive Digital Entertainment Conference*.
- [28] Sandra G Hart. 2006. NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 50. Sage publications Sage CA: Los Angeles, CA, 904–908.
- [29] Dorien Herremans, Ching-Hua Chuan, and Elaine Chew. 2017. A Functional Taxonomy of Music Generation Systems. *ACM Comput. Surv.* 50, 5, Article 69 (sep 2017), 30 pages. <https://doi.org/10.1145/3108242>
- [30] Ari Holtzman, Jan Buys, Maxwell Forbes, Antoine Bosselut, David Golub, and Yejin Choi. 2018. Learning to write with cooperative discriminators. *arXiv preprint arXiv:1805.06087* (2018).
- [31] Min-Chai Hsieh and Hao-Chiang Koong Lin. 2006. Interaction design based on augmented reality technologies for English vocabulary learning. In *Proceedings of the 18th International Conference on Computers in Education*, Vol. 1. 663–666.
- [32] Yueh-Min Huang, Yong-Ming Huang, Shu-Hsien Huang, and Yen-Ting Lin. 2012. A ubiquitous English vocabulary learning system: Evidence of active/passive attitudes vs. usefulness/ease-of-use. *Computers & Education* 58, 1 (2012), 273–282.
- [33] Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics* 8 (2020), 423–438.
- [34] Eunhyung Jo, Daniel A. Epstein, Hyunhoon Jung, and Young-Ho Kim. 2023. Understanding the Benefits and Challenges of Deploying Conversational AI Leveraging Large Language Models for Public Health Intervention. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 18, 16 pages. <https://doi.org/10.1145/3544548.3581503>
- [35] Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858* (2019).
- [36] 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.
- [37] Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. 2020. Gedi: Generative discriminator guided sequence generation. *arXiv preprint arXiv:2009.06367* (2020).
- [38] Mirella Lapata, Regina Barzilay, et al. 2005. Automatic evaluation of text coherence: Models and representations. In *IJCAI*, Vol. 5. Citeseer, 1085–1090.
- [39] Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691* (2021).
- [40] Jiwei Li, Will Monroe, and Dan Jurafsky. 2017. Learning to decode for future success. *arXiv preprint arXiv:1701.06549* (2017).
- [41] Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190* (2021).
- [42] Yitong Li, Martin Min, Dinghan Shen, David Carlson, and Lawrence Carin. 2018. Video generation from text. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.
- [43] Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021. GPT understands, too. *arXiv preprint arXiv:2103.10385* (2021).
- [44] Lara Martin, Prithviraj Ammanabrolu, Xinyu Wang, William Hancock, Shruti Singh, Brent Harrison, and Mark Riedl. 2018. Event representations for automated story generation with deep neural nets. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.
- [45] Piotr Mirowski, Kory W. Mathewson, Jaylen Pittman, and Richard Evans. 2023. Co-Writing Screenplays and Theatre Scripts with Language Models: Evaluation by Industry Professionals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 355, 34 pages. <https://doi.org/10.1145/3544548.3581225>
- [46] Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A Corpus and Cloze Evaluation for Deeper Understanding of Commonsense Stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for

- Computational Linguistics, San Diego, California, 839–849. <https://doi.org/10.18653/v1/N16-1098>
- [47] Paul Nation. 2007. The four strands. *International Journal of Innovation in Language Learning and Teaching* 1, 1 (2007), 2–13.
- [48] Paul Nation and Teresa Chung. 2009. Teaching and testing vocabulary. *The handbook of language teaching* (2009), 543–559.
- [49] Aurélien Nioche, Pierre-Alexandre Murena, Carlos de la Torre-Ortiz, and Antti Oulasvirta. 2021. Improving artificial teachers by considering how people learn and forget. In *26th International Conference on Intelligent User Interfaces*. 445–453.
- [50] Aurélien Nioche, Pierre-Alexandre Murena, Carlos de la Torre-Ortiz, and Antti Oulasvirta. 2021. Improving Artificial Teachers by Considering How People Learn and Forget. In *26th International Conference on Intelligent User Interfaces* (College Station, TX, USA) (IUI '21). Association for Computing Machinery, New York, NY, USA, 445–453. <https://doi.org/10.1145/3397481.3450696>
- [51] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]
- [52] Rebecca Oxford and David Crookall. 1990. Vocabulary learning: A critical analysis of techniques. *TESL Canada Journal* (1990), 09–30.
- [53] Rebecca L Oxford, Roberta Z Lavine, and David Crookall. 1989. Language learning strategies, the communicative approach, and their classroom implications. *Foreign Language Annals* 22, 1 (1989), 29–39.
- [54] 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.
- [55] Nanyun Peng, Marjan Ghazvininejad, Jonathan May, and Kevin Knight. 2018. Towards controllable story generation. In *Proceedings of the First Workshop on Storytelling*. 43–49.
- [56] Zhenhui Peng, Qingyu Guo, Ka Wing Tsang, and Xiaojuan Ma. 2020. Exploring the Effects of Technological Writing Assistance for Support Providers in Online Mental Health Community. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '20). Association for Computing Machinery, New York, NY, USA, 1–15. <https://doi.org/10.1145/3313831.3376695>
- [57] Savvas Petridis, Nicholas Diakopoulos, Kevin Crowston, Mark Hansen, Keren Henderson, Stan Jastrzebski, Jeffrey V Nickerson, and Lydia B Chilton. 2023. AngleKindling: Supporting Journalistic Angle Ideation with Large Language Models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 225, 16 pages. <https://doi.org/10.1145/3544548.3580907>
- [58] Julie Porteous and Marc Cavazza. 2009. Controlling narrative generation with planning trajectories: the role of constraints. In *Joint International Conference on Interactive Digital Storytelling*. Springer, 234–245.
- [59] Michael Pressley, Joel R Levin, and Harold D Delaney. 1982. The mnemonic keyword method. *Review of Educational Research* 52, 1 (1982), 61–91.
- [60] Mei Pu and Zheng Zhong. 2018. Development of a situational interaction game for improving preschool children's performance in English-vocabulary learning. In *Proceedings of the 2018 international conference on distance education and learning*. 88–92.
- [61] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.
- [62] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683* (2019).
- [63] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084* (2019).
- [64] Mark O Riedl and Robert Michael Young. 2010. Narrative planning: Balancing plot and character. *Journal of Artificial Intelligence Research* 39 (2010), 217–268.
- [65] Melissa Roemmele, Andrew S Gordon, and Reid Swanson. 2017. Evaluating story generation systems using automated linguistic analyses. In *SIGKDD 2017 Workshop on Machine Learning for Creativity*. 13–17.
- [66] Sherry Ruan, Liwei Jiang, Qianyao Xu, Zhiyuan Liu, Glenn M Davis, Emma Brunskill, and James A Landay. 2021. Englishbot: An ai-powered conversational system for second language learning. In *26th international conference on intelligent user interfaces*. 434–444.
- [67] Nathan Sakunkoo and Pattie Sakunkoo. 2013. Gliflix: Using movie subtitles for language learning. In *Proceedings of the 26th Symposium on User Interface Software and Technology*. ACM.
- [68] Marc Ericson C Santos, Takafumi Taketomi, Goshiro Yamamoto, Ma Rodrigo, T Mercedes, Christian Sandor, Hirokazu Kato, et al. 2016. Augmented reality as multimedia: the case for situated vocabulary learning. *Research and Practice in Technology Enhanced Learning* 11, 1 (2016), 1–23.
- [69] Norbert Schmitt. 2008. Instructed second language vocabulary learning. *Language teaching research* 12, 3 (2008), 329–363.
- [70] Abigail See, Aneesh Pappu, Rohun Saxena, Akhila Yerukola, and Christopher D Manning. 2019. Do massively pretrained language models make better storytellers? *arXiv preprint arXiv:1909.10705* (2019).
- [71] Yizhan Shao, Tong Shao, Minghao Wang, Peng Wang, and Jie Gao. 2021. A sentiment and style controllable approach for chinese poetry generation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 4784–4788.
- [72] Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980* (2020).
- [73] Sangho Suh, Jian Zhao, and Edith Law. 2022. Codetoon: Story ideation, auto comic generation, and structure mapping for code-driven storytelling. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–16.
- [74] Janet K Swaffar. 1988. Readers, texts, and second languages: The interactive processes. *The Modern Language Journal* 72, 2 (1988), 123–149.
- [75] Višnja Pavii Taka. 2008. *Vocabulary learning strategies and foreign language acquisition*. Multilingual Matters.
- [76] Preecha Tangworakitthaworn, Preeyapol Owatsuwan, Nutsima Nongyai, and Nongnapas Arayapong. 2019. An Image-Based Vocabulary Learning System Based on Multi-Agent System. In *2019 16th International Joint Conference on Computer Science and Software Engineering (JCSSE)*. IEEE, 324–329.
- [77] Scott R Turner. 2014. *The creative process: A computer model of storytelling and creativity*. Psychology Press.
- [78] Tim Van de Cruys. 2020. Automatic poetry generation from prosaic text. In *Proceedings of the 58th annual meeting of the association for computational linguistics*. 2471–2480.
- [79] Viswanath Venkatesh and Hillol Bala. 2008. Technology Acceptance Model 3 and a Research Agenda on Interventions. *Decision Sciences* 39, 2 (2008), 273–315. <https://doi.org/10.1111/j.1540-5915.2008.00192.x> arXiv:<https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1540-5915.2008.00192.x>
- [80] Thiemo Wambsgans, Tobias Kueng, Matthias Soellner, and Jan Marco Leimeister. 2021. ArgueTutor: An Adaptive Dialog-Based Learning System for Argumentation Skills. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) (CHI '21). Association for Computing Machinery, New York, NY, USA, Article 683, 13 pages. <https://doi.org/10.1145/3411764.3445781>
- [81] Thiemo Wambsgans, Christina Niklas, Matthias Cetto, Matthias Söllner, Siegfried Handschuh, and Jan Marco Leimeister. 2020. AL: An Adaptive Learning Support System for Argumentation Skills. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '20). Association for Computing Machinery, New York, NY, USA, 1–14. <https://doi.org/10.1145/3313831.3376732>
- [82] Lu Wang, Munif Ishad Mujib, Jake Williams, George Demiris, and Jina Huh-Yoo. 2021. An Evaluation of Generative Pre-Training Model-based Therapy Chatbot for Caregivers. *arXiv preprint arXiv:2107.13115* (2021).
- [83] Xingbo Wang, Haipeng Zeng, Yong Wang, Aoyu Wu, Zhida Sun, Xiaojuan Ma, and Huamin Qu. 2020. Voicecoach: Interactive evidence-based training for voice modulation skills in public speaking. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.
- [84] Stephen G Ware, R Michael Young, Brent Harrison, and David L Roberts. 2013. A computational model of plan-based narrative conflict at the fabula level. *IEEE Transactions on Computational Intelligence and AI in Games* 6, 3 (2013), 271–288.
- [85] Ziming Wu, Yulun Jiang, Yiding Liu, and Xiaojuan Ma. 2020. *Predicting and Diagnosing User Engagement with Mobile UI Animation via a Data-Driven Approach*. Association for Computing Machinery, New York, NY, USA, 1–13. <https://doi.org/10.1145/3313831.3376324>
- [86] Meng Xia, Qian Zhu, Xingbo Wang, Fei Nie, Huamin Qu, and Xiaojuan Ma. 2022. Persua: A visual interactive system to enhance the persuasiveness of arguments in online discussion. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–30.
- [87] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. 2018. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1316–1324.
- [88] Keiko Yamamoto, Jesus Rodriguez, and Yoshihiro Tsujino. 2019. FinDo: A Foreign Language Vocabulary Learning System Based on Location-Context. In *2019 20th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*. IEEE, 302–307.
- [89] Lili Yao, Nanyun Peng, Ralph Weischedel, Kevin Knight, Dongyan Zhao, and Rui Yan. 2019. Plan-and-write: Towards better automatic storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 7378–7385.
- [90] Liren Zeng and Ling Lin. 2011. An interactive vocabulary learning system based on word frequency lists and Ebbinghaus' curve of forgetting. In *2011 Workshop on Digital Media and Digital Content Management*. IEEE, 313–317.
- [91] Chao Zhang, Cheng Yao, Jiayi Wu, Weijia Lin, Lijuan Liu, Ge Yan, and Fangtian Ying. 2022. StoryDrawer: A Child-AI Collaborative Drawing System to Support Children's Creative Visual Storytelling. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–15.
- [92] Zheng Zhang, Ying Xu, Yanhao Wang, Bingsheng Yao, Daniel Ritchie, Tongshuang Wu, Mo Yu, Dakuo Wang, and Toby Jia-Jun Li. 2022. Storybuddy: A

human-ai collaborative chatbot for parent-child interactive storytelling with flexible parental involvement. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–21.