Open Sesame? Open Salami! Personalizing Vocabulary Assessment-Intervention for Children via Pervasive Profiling and Bespoke Storybook Generation

Jungeun Lee
jelee@postech.ac.kr
Dept. of CSE
POSTECH
Pohang, South Korea

Suwon Yoon
syoon25@postech.ac.kr
Dept. of CSE
POSTECH
Pohang, South Korea

Kyoosik Lee
kyoosik.lee@postech.ac.kr
Dept. of CSE
POSTECH
Pohang, South Korea

Eunae Jeong
eunaec@ewhain.net
Dept. of Communication Disorders
Ewha Womans University
Seoul, South Korea

Jae-Eun Cho
jaeeuncho@ewhain.net
Dept. of Communication Disorders
Ewha Womans University
Seoul, South Korea

Wonjeong Park
thisisforwj@gmail.com
Dept. of Communication Disorders
Ewha Womans University
Seoul, South Korea

Dongsun Yim
sunyim@ewha.ac.kr
Dept. of Communication Disorders
Ewha Womans University
Seoul, South Korea

Inseok Hwang
i.hwang@postech.ac.kr
Dept. of CSE
POSTECH
Pohang, South Korea

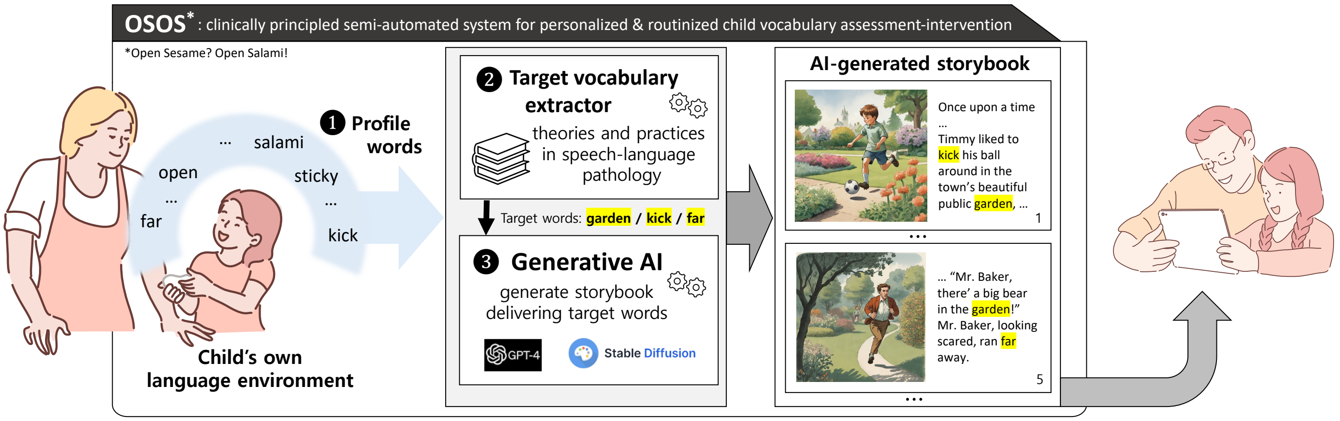

OSOS*: clinically principled semi-automated system for personalized & routinized child vocabulary assessment-intervention
*Open Sesame? Open Salami!
① Profile words
... salami
open
sticky
...
...
far
kick
Child's own language environment
② Target vocabulary extractor
theories and practices in speech-language pathology
Target words: garden / kick / far
③ Generative AI
generate storybook delivering target words
GPT-4
Stable Diffusion
AI-generated storybook
Once upon a time ... Timmy liked to kick his ball around in the town's beautiful public garden, ...
1
... "Mr. Baker, there' a big bear in the garden!" Mr. Baker, looking scared, ran far away.
5

Figure 1: OSOS, a pervasive, generative, and clinically principled system for personalized vocabulary assessment and intervention for children in everyday routines

ABSTRACT

Children acquire language by interacting with their surroundings. Due to the different language environments each child is exposed to, the words they encounter and need in their life vary. Despite the standard tools for assessment and intervention as per predefined vocabulary sets, speech-language pathologists and parents struggle with the absence of systematic tools for child-specific custom vocabulary, i.e., out-of-standard but personally more important. We propose "Open Sesame? Open Salami! (OSOS)", a personalized vocabulary assessment and intervention system with pervasive language profiling and targeted storybook generation, collaboratively developed with speech-language pathologists. Melded into a child's daily life and powered by large language models (LLM), OSOS profiles the child's language environment, extracts priority words therein, and generates bespoke storybooks naturally incorporating those words. We evaluated OSOS through 4-week-long deployments to 9 families. We report their experiences with OSOS, and its implications in supporting personalization outside standards.

CCS CONCEPTS

• **Human-centered computing** → *Ubiquitous and mobile computing systems and tools*; • **Social and professional topics** → *Children*; • **Applied computing** → *Health care information systems.*

KEYWORDS

language assessment and intervention, vocabulary learning, storybook generation, generative AI, large language model

---

Permission to make digital or hard copies of part or all 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 third-party components of this work must be honored. For all other uses, contact the owner/author(s).

CHI '24, May 11–16, 2024, Honolulu, HI, USA
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0330-0/24/05.
https://doi.org/10.1145/3613904.3642580

ACM Reference Format:
Jungeun Lee, Suwon Yoon, Kyoosik Lee, Eunae Jeong, Jae-Eun Cho, Wonjeong Park, Dongsun Yim, and Inseok Hwang. 2024. Open Sesame? Open Salami! Personalizing Vocabulary Assessment-Intervention for Children via Pervasive Profiling and Bespoke Storybook Generation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 32 pages. https://doi.org/10.1145/3613904.3642580

1 INTRODUCTION

Every child is different. So is their surrounding language environment. Which culture and era they live in, whom they frequently interact with, and even their family's socioeconomic status contribute to making a child's language environment distinct [73, 117]. Such diversity influences the type and range of *vocabulary* a child encounters in their daily interactions [17], which is the most basic linguistic element that develops from the earliest age and broadly impacts later language development.

A preschooler's vocabulary is highly representative of academic performance in school age [55]. Vocabulary assessment in these ages is accepted as a measure of a possible language delay [57], which is correlated to further aggravation such as learning disorders [75], psychiatric disorders [26], or even low socioeconomic status [43] unless timely assessed and intervened. Standardized tools for child vocabulary assessment (e.g., M-B CDI [57], PPVT [51]) define a scoring system based on a small set of standard words sampled from large linguistic studies. However, standardized tools pose multiple limitations. It is extremely costly to revise; e.g., 16 years between revisions [57, 125], unchanged over a decade [139]. Naturally, it is rigid to the diachronic evolution of language [62]; a 2007 tool has no word about smartphones which are now everywhere [57]. Consequently and most importantly, it is unable to personalize. Despite the diversity of language environment each child is exposed to in their daily life and hence the nonuniform vocabulary they naturally develop, the uniform set of words that standardized tools test makes it difficult to evaluate a child's actual language abilities in personalized terms [143].

Interventions also face personalization challenges. Speech-language pathologists (SLPs) determine the target words for a child, and facilitate their learning through intervention aids, e.g., pictures or books. The clinical setup is detached from the child's daily routines; the SLPs lack direct and comprehensive access to one's language environment. The SLP's interventions are inevitably spotty. Furthermore, ready-made intervention aids often do not cover various words that matter to the child. Overall, given the recent individualization of children's language environments, we find SLPs increasingly seek methodologies to help child vocabulary assessment and intervention be *personalized* and *routinized.*

We present "*Open Sesame? Open Salami!*" (*OSOS*) – a pervasive, generative, and clinically principled system to enable personalized vocabulary assessment and intervention, by blending into a child's everyday routines and bespoke creation of intervention aids. Specifically, OSOS is designed to work on smart devices at home, to (1) unobtrusively profile a child's language environment, (2) distill personally prioritized vocabulary for intervention, and (3) generate personalized storybooks to naturally nourish the child with their personalized target words in daily routines. OSOS is built upon LLMs (large language models), generative visual models, and commodity mobile devices.

Developed by an interdisciplinary team of SLPs and computer scientists, OSOS deeply incorporates clinical principles and practices into its design with computational feasibility. In monitoring daily verbalizations, OSOS profiles the vocabulary therein as per linguistic parameters [23] influential to childhood vocabulary acquisition. In extracting personalized priority words, OSOS is of a modular architecture allowing different prioritizing criteria, reflecting the multi-ended nature of developmental goals. Currently, OSOS is prototyped with the occurrence-based model [22] (an exemplary set of criteria widely exercised in clinical practices utilizing a word's frequency, perceptual salience, commonality, etc. [111]; detailed in §5.2), while being open to other goal-specific criteria [25, 70, 132]. OSOS's design rationale of utilizing storybooks as routinizable intervention aids is consistent with the child development theories [130, 169, 174, 193] and common familial routines [146, 147]. In generating personalized storybooks incorporating the target words, OSOS provides a near-automated generative pipeline for textual and pictorial narratives along with a convenient human-in-the-loop (possibly by an SLP) web interface to expedite principled revision as per story grammar [27, 109].

OSOS excels in the level of personalization and routinization of the assessment-intervention cycle of child vocabulary. Long-time blended into a child's daily life, OSOS develops its personalized understanding and intervention as per a child's vocabulary needs in *comprehensive*, *generative*, and *embodied* manners, rather than momentary, curative, or instructive [65, 86, 108, 166]. OSOS also adds a novel exploration of principled storybook generation grounded on pathological frameworks for child vocabulary intervention, onto the existing literature of storybook generations [31, 61, 67, 71, 115, 192, 194] and human-AI story co-creation [41, 137, 160, 191].

We designed, developed, and deployed OSOS over three phases. We conducted a preliminary study with 4 licensed SLPs, finding hands-on lessons and challenges. Then, we, 4 computer scientists and 4 SLPs, iteratively designed and developed OSOS over 5 months period. OSOS integrates Android tablets, GPT-4 [136], Stable Diffusion XL [145], and speech/linguistic libraries [4, 7]. Finally, we deployed OSOS for 4 weeks onto 9 families with preschool children, where OSOS produced a total of 120 personalized target words and 180 AI-generated bespoke storybooks with target words embedded. We investigated the clinical implications and user experiences.

Our contributions are threefold. (1) We explore the systematic feasibility of clinically principled personalization of child vocabulary assessment and intervention in daily context. (2) We premier an initial implementation of personalized vocabulary prioritization and incorporation with generated storybooks grounded on clinical frameworks. (3) We report comprehensive findings from a 4-week in-the-wild deployment and discuss multi-faceted implications.

We organize the paper as follows. §2 reviews the clinical backgrounds and HCI- & AI-driven works on child vocabulary intervention. §3 outlines our study procedure. §4 presents our preliminary studies with 4 SLPs. §5 depicts the architecture of OSOS and principled implementations. §6 and §7 present the deployment setup and findings, respectively. We discuss the implications and limitations in §8, before concluding the paper.

2 BACKGROUND AND RELATED WORK

2.1 Vocabulary Acquisition of Children

Vocabulary acquisition is a foundational process in early language development. A child acquires several protowords around 12 months and 50+ around 18 months. Eventually, they acquire 10,000+ words when entering the school age [18]. The vocabulary at this stage has long-term effects on their literacy skills and academic achievement [55], necessitating timely assessment and interventions that ensure enriched vocabulary environments.

Children's vocabulary development is influenced by internal factors – e.g., individual cognitive skills [141], and external factors – quantity and quality of language input [17]. Fast mapping [32] explains the mechanism of how young children learn vocabulary given few exposures to a word in a relevant context. As children learn vocabulary through external interactions, their surroundings significantly influence one's vocabulary. Notably, the digitization of daily life is diversifying the language environment, taking language input from various sources, e.g., on-demand videos, e-books, smart devices, etc. [54, 98, 197], where the collections and contents vary across families. Hence, children tend to learn certain vocabulary more extensively depending on individual environments.

Given language input, the acquisition of each word differs by its linguistic factors – frequency, redundancy, perceptual salience, etc. [22, 23]. Frequency represents how often a word is exposed to a child [68, 74, 78, 79]. Perceptual salience refers to how clearly a word is perceived, in terms of speed, accent, rhythms, etc. [144, 185].

The higher a word is scored in one or more factors in a child's language environment, the more likely the child acquires that word than others. Conversely, a child not acquiring high-scored words might be an indication to early-screen their internal vocabulary development. It is reported that the relative impact of these factors on children's vocabulary acquisition would vary across children [168].

2.2 Standardized Assessment Tools for Vocabulary Development

10-20% of the preschool population is diagnosed with language delay [81, 157]. Vocabulary development is particularly important for them because it progresses faster than syntax, pragmatics, and phonology. Primary caregivers should timely assess their child's vocabulary acquisition skills, so that early intervention follows.

In clinical settings, standardized and non-standardized tests are used to evaluate a child's vocabulary development. Major standardized tests include MacArthur Bates Communication Inventories (M-B CDI) [57] and Peabody Picture Vocabulary Test (PPVT) [51]. Both tests provide a *predefined list of standard words*. In essence, a child's vocabulary development is assessed by checking how many words she has acquired, out of the whole standard words. M-B CDI presents the standard words to the caregiver and, according to their memory, asks them to check off the words that the child knows. PPVT presents the standard words and a set of pictures directly to the child and elicits their responses.

A major limitation in standardized tests is to determine the child's vocabulary only within the standard words (e.g., 680 words in M-B CDI toddler-long form) [59]. Their adherence to the standard words leaves many non-standard real-life words not taken into account. Moreover, a standard word set is rigid; once defined, it remains unchanged for a long time. The latest M-B CDI update interval was 16 years [57, 125]. Updating a standard word set is extremely demanding – a multi-year task involving nationwide in-person surveys and diagnosis, followed by extensive post-analyses. These challenges make standard tests not up-to-date with contemporary words [117]. To be more complicated, individual children, even in the same developmental progress, exhibit a large diversity in the order of words that they learn earlier or later, attributable to individual differences in regional dialects [44, 126, 161], cultural ancestry [48, 80], socio-economic status [138], and family mood [127]. Standard tests are not considerate of such individual differences despite all being in normal vocabulary development, likely resulting in over- or under-estimation of one or another. Given the aforementioned limitations and also increasing differences between children's language environments (§2.1), real-life words may matter more to fairly evaluate a child's actual language abilities [143].

To complement, non-standardized approaches are conducted in clinics [175, 188]. Clinicians present a situation where the child interacts with an adult, and the child's utterances are analyzed [190]. Clinicians consider non-standardized tests essential to assess a child's actual language development [95, 143]. However, it requires one-on-one assessment by a trained clinician, making it highly costly and little affordable to many families [152, 179].

2.3 Vocabulary Intervention for Children

Among many intervention activities designed to foster a child's vocabulary skills, book-reading with an adult is well-known to be effective in building preschool children's vocabulary [134, 159, 181, 183]. Picture books expose the child to the linguistic context with natural combinations of pictures and text, enriching vocabulary development [130, 169, 174, 193]. Picture book reading is a universal activity in almost every home [197], making it a natural choice of routine activities for various augmentative purposes in HCI communities [93, 146, 147, 195].

Clinicians frequently use book-reading to teach specific target words. However, it is difficult to select books that contain the non-standard target words chosen for each child. Moreover, the clinicians plan a series of words as per-child assessment goals over a period, thereby needing a series of intervention aids in line with the series of words planned ahead [143]. Furthermore, clinical practices recommend multiple books for a target word. Therefore, the difficulty of arranging books in line with the intervention plan multiplies with child-specific target vocabulary.

2.4 Computational Tools for Child Language Development & Intervention

HCI communities have developed pervasive tracking of children's development. Baby Steps [96] tracks one's development through in-situ recordings, later incorporating Twitter [171] and SMS [172]. BebeCODE [165] presents a mobile system for agreeable assessment between caregivers. Toys are retrofitted with sensors for life-immersive tracking [34, 89, 182], facilitating parent-driven screenings along real-life. For language-specific tracking, however, these systems are based on predefined questions, inflexible to different language environments around individual children.

LENA [65] is a commercial tool for a child's out-of-clinic speech collection and assessment. It features a pocket-sized recorder (over $300) and cloud services. LENA also produces reports of the child's speech duration, turn-taking, etc. so that the parents compare their child's metrics to the standard milestones. In essence, LENA helps data-driven assessment while leaving interventions to SLPs or parents. OSOS features interventions that are (1) embodied, not instructive; (2) personally generative, not curative; (3) aiming at individually diverse goals, rather than standard ones.

HCI communities have developed real-time language intervention systems on mobile devices. TalkBetter [85, 86] monitors the structural progress of parent-child conversations and issues real-time reminders. TalkLIME [166] reinforces parent training through non-invasive goal-driven interventions. Captivate! [108] displays context-specific word cards for immigrant parents. While these systems excel in real-time interventions, they are limited to standard interventions curated upon a momentary context.

Overall, OSOS stands out as a clinically-inspired pervasive system specializing in personalized child vocabulary assessment and intervention. OSOS continuously profiles the language to & from the child, tailors personally prioritized vocabulary goals, and generates intervention aids embodied in natural routines. OSOS would undergo a virtuous loop of personalized intervention resulting in updated assessment, in turn generating newer intervention, and so on. Broadly, OSOS shares the emerging theme calling for bespoke-generation of personalized norms out of conventional common ones, while being grounded on underlying principles and respecting the circumstantial diversity across individuals [113, 114].

2.5 Storytelling with Interactive Technologies

An ample body of works presented novel designs and studies on children's storytelling combined with digital technology. 'This book is magical!' [177] explored the engagement of children in storytelling with traditional books and various digital forms including interactive books, robots, etc. ROBIN [176] presented a design prototype for a humanoid robot to engage children in narrative-building. 'Fiabot!' [153] demonstrated the practicality of digital support on the school curriculum while giving design guides to stimulate creativity in children's story authoring [154]. Kids in Fairytale [94] utilized Mixed Reality for children's immersive storytelling.

Facilitating storybook authoring by children has been actively explored. Our Story [104], 1001stories [50], and Rubegni et al. [155] involved children as active creators, maximizing their sense of agency [104] and engagement [50], and guiding the design of digital storytelling authoring tools for preliterate children's pre-reading and writing [155]. Inspired by these works, we envision OSOS may retrofit its story steering console (depicted in Figure 7 and §5.3) to help children steer the content of the generated storybooks. Note that OSOS provides semi-automatic generation of new personalized storybooks, while the above works largely utilize existing storybooks and/or manual creation.

Embedding personalized elements in storytelling showed immersion and creativity. People in Books [61] embeds family members' faces into storybooks. Zaturi [93] microtasks the parents to create audiobooks in their own voice. TellTable [31] creates stories with physical elements around the child. Zarei et al. [192] lets children embody an avatar as the story character. Child-AI collaborative systems facilitate creative storytelling [71, 194] and rewriting [115].

Kucirkova et al. have presented extensive studies on personalized storytelling in the child education context. A series of works [105–107] incorporated personalized elements such as the child's name, photo, favorite toy, etc. into storybooks. They investigated the child's acquisition of pre-determined target words [105, 106] and parent-child dyad's engagement [107]. OSOS explores an orthogonal dimension for story personalization: each child's daily language environments and personally important words therein. Due to the higher diversity and time-variation of everyday words compared to children's names etc., OSOS employs generative AIs to produce new stories that blend with diverse target words. Overall, we believe that the aforementioned works and OSOS may be combined to make storytelling experiences even richer, more immersive, and closely aligned to one's everyday context.

2.6 Automating Storybook Generation

Computational (semi-)automation of story generation, which dates back to the 1960s [16, 156], has flourished along with the advance of collective intelligence and generative AI. Crowdsourcing facilitated domain modeling [118], asymmetric contribution [99], goal decomposition [100], and role-play [82] for systematic acceleration of story-writing. In neural story generation, the sequential properties of stories made recurrent neural networks (RNN) an early choice [40, 72]. Hierarchical story generators addressed the limited event sequences of RNNs and improved coherence [56, 186]. The emergence of large language models (LLMs) facilitated human-LLM co-creation [137, 160]. Wordcraft [191] is a text editor with LLM-powered rephrasing or continuation. TaleBrush [41] enables control over the protagonist's fortune. PlotMachines [150] presents outline-conditioned story generation.

For children's storybooks, it is imperative to pair the text with images. Although major generation models [10, 145, 149] produce high-quality images reflecting the prompt, storybook images need engineering beyond discrete text-to-image. StoryDALL-E [122] visualizes long metaphorical narratives rather than a brief caption. Make-a-Story [148] employs visual memory modules to keep actor- and scene-consistency. TaleCrafter [67] visualizes interactive stories featuring selective editability on layouts and structures.

As discussed, computer-assisted storybook generation has been studied at various degrees of automation and human-in-the-loop. Latest incorporation of LLMs and generative AIs enabled *on-the-fly* generation of custom-steered storybooks at practical quality and scalability. Despite a series of works on (semi-)automatic generation of children's storybooks, to the best of our knowledge, principled storybook generation grounded on the speech-language pathological frameworks is little explored. In this light, we present an initial work that leverages AI-generated personalized storybooks as a targeted intervention aid for child-specific vocabulary learning. Furthermore, we put this into a larger system of assessment-intervention loop that naturally blends within the child's own book-reading routines.

For ethical precaution in deploying an experimental system onto children, we purposefully refrained from full-automation. OSOS is largely automated but has a few human-in-the-loop screenings.

We acknowledge that our initial work leaves room to improve its human-AI co-creation interfaces, story-specific generative models, and degree of automation. These are next-step quests, on which this paper would shed light for informed exploration.

3 STUDY PROCEDURE

Our background studies in §2 lead us to three key motivations:

(1) **Personalizing a child's vocabulary assessment** (from §2.1 and §2.2). Despite the importance of timely assessing a child's vocabulary [33, 55], the prohibitive cost of building or updating standard vocabulary [57, 117, 125] mandates one-size-fits-all assessment, regardless of individual diversity of language environments. For example, 'teddy bear' is a standard word of M-B CDI [57]. If a toddler does not own a teddy bear, she might not know it, but if she often uses her parent's iPhone, that word is probably in her daily vocabulary. Standard tools ignore her knowledge of 'iPhone' while penalizing her for not knowing 'teddy bear'. iPhone's socio-economic bias [72] may limit its use as a standard word. Yet, iPhone effectively indicates her acquisition of prevalent words *in her environment.* To fairly assess each child's vocabulary learning ability, we call for a personalized tool that complements standard tools by accounting for the child's unique vocabulary exposure.

(2) **Personalizing a child's vocabulary intervention** (from §2.2 and §2.3). Assessing a child's vocabulary often derives target words that this child needs to learn [95, 175, 188, 190]. Given today's vocabulary intervention often done with ready-made aids [143, 181, 193], personalized vocabulary assessment will naturally call for personalizable intervention aids, so that an unbounded set of arbitrary words can be supported.

(3) **Blending the assessment-intervention process into the child's daily routines** (from §2.2 through §2.4). To complement today's clinical practices [51, 57] and facilitate personalization, we call for blending personalized vocabulary assessment and intervention into an individual child's natural routines. This is in line with various HCI & UbiComp efforts pursuing an everyday commodity tool for early-screening [47, 110, 180] and routinized intervention [86, 91, 102, 189] to broaden clinically implicative feedback into real life.

In this light, we envision a pervasive system that helps vocabulary assessment-intervention in a *personalized* and *routinized* manner. To validate, develop, and evaluate this vision, we conducted three phases of study (Figure 2).

Phase 1: Preliminary study with domain experts. We consulted 4 licensed SLPs through 1-on-1 interviews. We learned about their clinical practices of vocabulary assessment and intervention, including the rigid uniformity of today's standard tools. §4 presents the key findings.

Phase 2: Iterative design and development of our system, *OSOS*. We, an interdisciplinary team of computer scientists and speech-language experts, iteratively refined the design of OSOS as per clinical practices, computational feasibility, and usability. OSOS is carefully engineered to the quality for home deployment. OSOS continuously profiles daily conversations around the child, distills customized target words for intervention, and generates personalized storybooks naturally embedding the target words into brand-new stories. OSOS is built with tablets, speech-to-text (STT) models [4], LLMs [136], and image generation models [145]. §5 describes the architecture and design rationales.

Phase 3: In-the-wild deployment study. We deployed OSOS to 9 families with children of our target ages (4-5 years), lasting 4 weeks per family. The earlier 2 weeks are to profile personalized language environments and to extract child-specific target words. For the later 2 weeks, each family was provided 20 OSOS-generated storybooks such that one's own target words are naturally embedded. §6 explains our deployment setup. §7 reports the findings.

Before detailing our study, we clarify that our personalized approach is not to replace the existing standard tools, but to complement them. For horizontal comparisons between children, standard tools serve as an agreeable common ground. In contrast, we address an intra-child perspective – to understand how a child is learning vocabulary with respect to her own circumstances, identify which words this child would need to learn, and provide personally tailored intervention, through an individually dedicated lens rather than comparison against others. This perspective has been underexplored by current standard tools where individual circumstances are inevitably approximated to a uniform one and many child-specific factors correlated to their vocabulary learning are neglected.

4 PRELIMINARY STUDY WITH EXPERTS

We report our findings from interviews with 4 licensed SLPs (denoted S1 through S4; 1.5 hours each). No author is an interviewee. The SLPs have an average career of 6.75 years (min: 4, max: 10) in child language assessment and intervention. We asked how they assess a child's vocabulary, how they plan a goal and intervention methods, and whether they experienced disparities between given tools and children. The interviews are transcribed and analyzed by 3 researchers. §4.1 through §4.5 report the high-level themes grounded on the data [170]. In §4.6, we summarize our design inspirations elicited from the interviews. All quotes are translated from the Korean language.

4.1 Paramount Importance of Childhood Vocabulary Development

All SLPs emphasized vocabulary in early language development. S1: "*Every language-delay case involves vocabulary issues.*" S4: "*For the 3-5 years old, vocabulary skills are a representative measure (of overall language skills).*" S1: "*The weight of vocabulary interventions does not diminish as they grow; rather it always go with other interventions.*"

4.2 Limitations in Current Language Assessment Methods

All SLPs stressed that the standardized methods get outdated quickly. For PPVT [51], the SLPs took many examples of outdated picture cards, e.g., a phone booth, a rotary dial phone, and a computer with a CRT monitor that today's children do not know. In contrast, the non-standard word 'administer' became a de-facto standard word known to almost every child, after COVID-19.

They evaluated that standardized methods may not fairly capture a child's actual vocabulary level. S4: "*Some children remember the standard words from previous tests, resulting in overestimation.*" S1

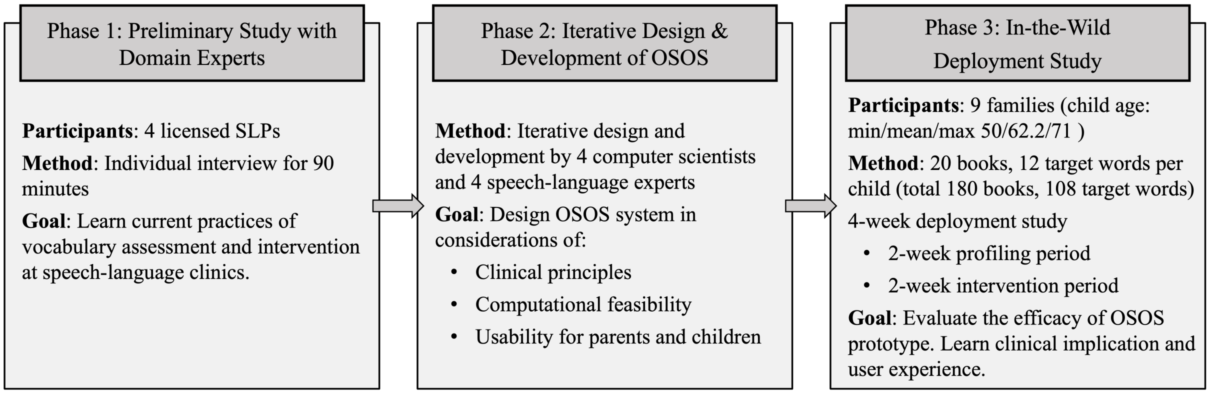

Phase 1: Preliminary Study with Domain Experts
Participants: 4 licensed SLPs
Method: Individual interview for 90 minutes
Goal: Learn current practices of vocabulary assessment and intervention at speech-language clinics.
Phase 2: Iterative Design & Development of OSOS
Method: Iterative design and development by 4 computer scientists and 4 speech-language experts
Goal: Design OSOS system in considerations of:
• Clinical principles
• Computational feasibility
• Usability for parents and children
Phase 3: In-the-Wild Deployment Study
Participants: 9 families (child age: min/mean/max 50/62.2/71)
Method: 20 books, 12 target words per child (total 180 books, 108 target words)
4-week deployment study
• 2-week profiling period
• 2-week intervention period
Goal: Evaluate the efficacy of OSOS prototype. Learn clinical implication and user experience.

Figure 2: 3-phased study procedure.

and S2 state that the unfamiliar test environment leads to underestimation. To complement, there are non-standardized methods, e.g., monitoring the child's self-verbalizations in an emulated play. However, it is highly burdensome. S1: "*Analyzing the monitored results is super labor-intensive. (...) Many clinics do only the standardized tests for cost reasons.*"

4.3 Limitations in Selecting *Which Words* To Teach

To select child-specific target words for intervention, SLPs refer to the test results and parent surveys. S1: "*The test results may indicate a category of words that the child needs.*" S2: "*I ask the parent about the child's daily life and the particular words that they'd need.*" S3: "*I ask for the words occurring in the preschool.*" They often put multiple relevant words together in a single session. S2: "*Towel, handkerchief.*" S3: "*Thread, needle.*"

The SLPs pointed out major limitations in target word selection – lack of reflection of a child's real-life language activities. S2: "*I'd like to look into their real-life conversations.*" S3: "*I want a tool that captures frequent words from the child's favorite YouTube.*" S4: "*I'd like feedback from the preschool where the child stays long.*" S1: "*Once I determine the target words, I want to prioritize them by the child's real-life need. But there's no reference.*"

4.4 Limitations in *How* to Teach the Words

Once the SLPs determine which words to teach, they seek intervention aids to convey those words. Play, book-reading, and picture cards are common options. We asked: "*Have you ever had an episode in which you could not find an interaction aid for the target word?*" S2: "*Very frequent.*" S1: "*Quite often.*" S4: "*3 times out of every 10.*" They elaborated: S1: "*I often make a new picture card myself, but it is too much work. I can't do every single word.*" The SLPs pointed out that interventions do not seamlessly propagate to home. S1: "*I provide parent instructions, but there's a large variance in how much they follow.*" S4: "*I used to create milestones of home activities. But they couldn't do that.*"

4.5 Vocabulary Intervention via Books

The SLPs unanimously appreciated book-reading with an adult, as an important, ubiquitous, and effective intervention aid. They valued that book-readings allow the words naturally exposed to various contexts. S4: "*Books are the best tool to teach vocabulary! The words appear with a context.* " S2: "*By reading books, they learn a word in more than one context.*" Meanwhile, they mentioned that book-readings are usually effective for 4+ years old.

The limitations of book-reading lie in their inflexibility to arbitrary words, similar to §4.4. S1: "*I can't do book-reading when I can't find a book with that word.*" To narrow the gap, SLPs often improvise. S3: "*I changed the characters to impersonate someone close to the child.*" S2: "*I mix up the target words and the child's experience, and make up a story.*" Albeit they alter stories to fit a child's needs, it is not applicable to the pictorial contents. S3: "*I was frustrated that I couldn't change the pictures.*"

Upon hypothetical technologies to dynamically change existing books, they were very positive. S2: "*Fascinating! That'll interest a child and let them learn better.*" They suggested ideas on how to change books. S2: "*It'd be awesome to change the context for a given word. Say 'Fly'. Then it's like 'Fly an airplane,' 'Fly a kite.'.*" S3: "*Put the target word into something that lively interacts in the story.*"

4.6 Design Inspirations from Expert Interviews

Our interviews inspired motivations and guidelines for building a novel system. We call for a system with the following features, listed along with the respective SLPs whom we identified relevant challenges and demands.

- An assessment method tailored to each child, to fairly reflect their actual vocabulary level. (S1, S2, S4; in §4.2)
- Target word selection reflecting an individual child's daily language environment. (all S1-S4; in §4.3)
- On-demand interaction aids that match an arbitrary target word. (S1, S2, S4; in §4.4)
- A family-friendly method for continual vocabulary intervention at home. (S1 and S4; in §4.4)
- A new breakthrough in book-reading for word- and child-adaptive vocabulary intervention. (S1, S2, S3; in §4.5)

5 OSOS: SYSTEM OVERVIEW

Figure 3 depicts the architecture of OSOS with three major modules: '*Personalized Language Profiler* (abbr. *Profiler*)', '*Target Vocabulary Extractor* (abbr. *Extractor*)', and '*Personalized Intervention Aid Generator* (abbr. *Generator*)'. §5.1 through §5.3 describe the design rationales and implementation of each module.

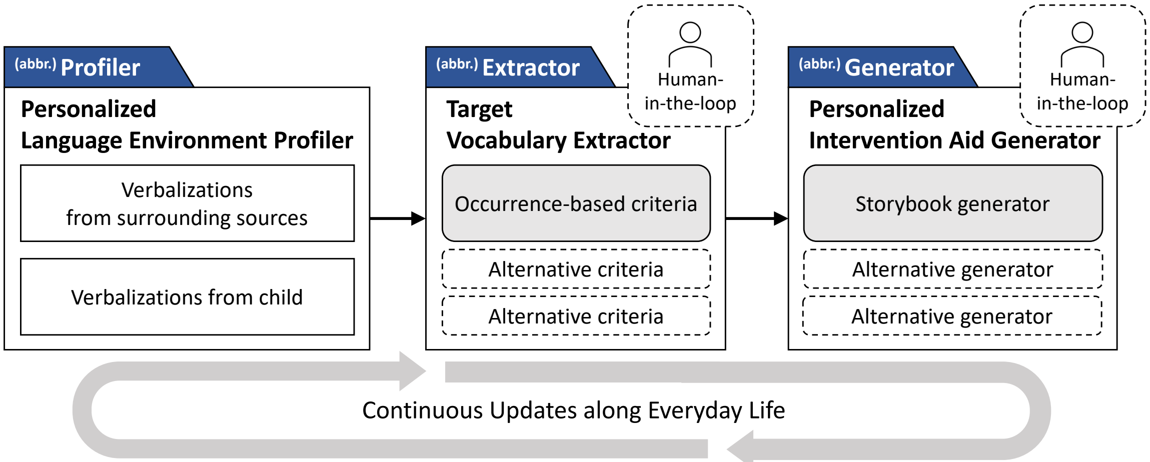

(abbr.) Profiler
Personalized Language Environment Profiler
Verbalizations from surrounding sources
Verbalizations from child
(abbr.) Extractor
Human-in-the-loop
Target Vocabulary Extractor
Occurrence-based criteria
Alternative criteria
Alternative criteria
(abbr.) Generator
Human-in-the-loop
Personalized Intervention Aid Generator
Storybook generator
Alternative generator
Alternative generator
Continuous Updates along Everyday Life

Figure 3: Overall architecture of OSOS.

Our studies in §2 and §4 led us to the initial sketch of a pervasive system that facilitates *personalized assessment and intervention* on a child's vocabulary, that blends into the child's own language environments along *their natural routines*.

To materialize this sketch into a concrete system, we formed an interdisciplinary team of four computer scientists specializing in HCI & ubiquitous computing and four speech-language experts (including two licensed SLPs) specializing in child language.

We targeted a real-working system of deployable quality, with selective human-in-the-loop for ethical precaution. Our development took 5+ months with a 2-day intensive on-site workshop, 3 times of in-person meetings, and weekly online meetings. We iteratively refined numerous design decisions and algorithms, by putting together the SLPs' domain knowledge, field practices, existing clinical datasets, and computational feasibility (e.g., speech-sensing accuracy, model-controllability, prompt engineering, quality of AI-generated content, and computing time for generations).

5.1 Personalized Language Environment Profiler (abbrev. Profiler)

The Profiler is intended to be deployed at the target child's home and to collect speech samples therein. The parent controls when to start and stop recording. We appropriated an Android smartphone as a programmable networked recorder. It performs as a stationary microphone itself, the base station for the wearable Bluetooth microphone, and an uploading agent. Figure 9 shows the wearable and stationary microphones. For productization, the Profiler may be embedded in smart speakers or home appliances, so that they collectively provide a larger coverage.

The software consists of mobile apps and backend servers. The mobile apps perform (1) voice-recording from both microphones, (2) parent-interface to control recording, and (3) uploading of parent-approved recordings. We built server software in Python that performs: (1) voice acquisition and management, and (2) automated speech-to-text (STT) transcription and speaker diarization based on CLOVA Speech [4].

5.2 Target Vocabulary Extractor (abbrev. Extractor)

The Extractor analyzes the utterances from the Profiler, and extracts a prioritized list of words recommendable for this child. The prioritization criteria are selectable. Literature teaches various criteria to prioritize the words that a child needs to learn next. Our default implementation is 'occurrence-based' – widely exercised criteria putting weight on the words that occur more frequently and commonly around the child, yet the child has not learned [22, 111].

Possible alternative criteria would be 'conceptual relevance' – putting weight on new words conceptually close to what the child knows [70, 132], or 'tier systems' – letting children self-learn high-frequency words and put weight on the words rarely-occurring but of high-utility [25]. The efficacy of certain criteria over another is ongoing research in the speech-language literature; comparisons between criteria are not the scope of our paper. We design our architecture to flexibly plug-and-play different criteria. For deployment, we adopt the 'occurrence-based' as they are widely exercised.

Figure 4 depicts the detailed flow of the Extractor implemented in Python on our local Linux server. The identifiers E_1, E_2, ..., E_6 indicate the corresponding locations in Figure 4.

E_1 For a given child, the Extractor takes the diarized transcriptions from the Profiler and divides them into two sets: the child's speech (S_C) and non-child speech (S_{NC}). POS (part-of-speech) analyzer [7] decomposes the transcriptions into words with POS labels. The words in S_{NC} are assumed the vocabulary that the child was exposed to. Several preprocessing steps are applied to S_C and S_{NC}. (1) Stopwords are filtered out. As a Korean stopword database [97] includes content words meaningful for OSOS, we did not use the database but filtered out function words by POS labels. This POS-based filtering may not be 100% inclusive, but we observed that the surviving stopwords are mostly filtered at E_5 (explained later) which keeps only the words belonging to $[S_{NC} - S_C]$, where S_C likely includes most of the surviving stopwords. (2) To mitigate possible STT errors, we excluded single-syllable words and the words that appeared only once during the two-week profiling. (3) For words of the same root but minor differences (e.g., suffix), only one form is selected. Note that our POS analyzer factorizes a word

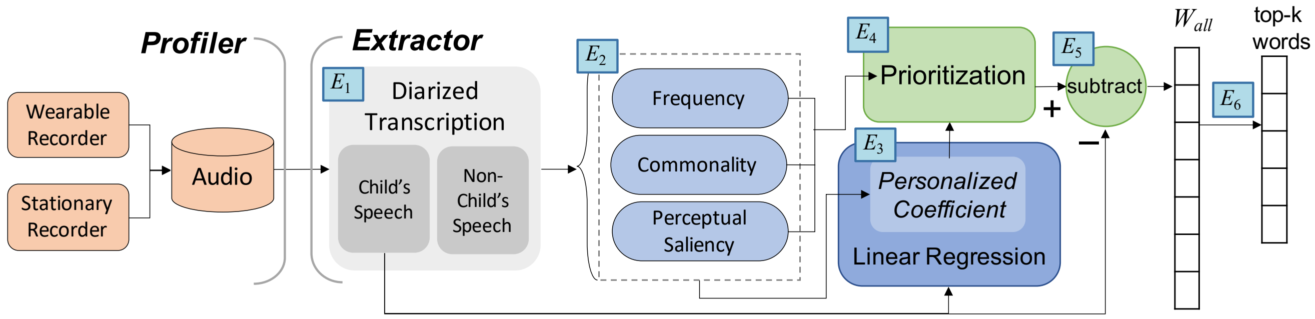


Figure 4: Detailed operation flow of Target Vocabulary Extractor (with implementation-choice of Occurrence-based criteria).

into morphemes where root-level identification is naturally done, due to the agglutinative properties of the Korean language [164].

$\boxed{E_2}$ For each unique word $w_i \in \mathrm{S}_{\mathrm{NC}}$, the Extractor evaluates its *frequency* $f(w_i)$, *commonality* $c(w_i)$, and *perceptual salience* $p(w_i)$ – the major parameters how the occurrence-based criteria determine the priority of w_i for that child [22, 23, 111]. The frequency refers to the number of occurrences of w_i in S_{NC}. The commonality refers to the number of distinct sources (e.g., days, places, speakers) that w_i appears from. The perceptual salience refers to how clearly w_i is articulated. The rationale of occurrence-based criteria is as follows. The higher frequency, commonality, and perceptual salience w_i exhibits, the more chances that the child would have learned w_i. Conversely, having not learned w_i implies that w_i is a priority word that this child should learn sooner.

$\boxed{E_3}$ Given the rationale and parameters to determine a word's relative priority, we need a quantitative expression to evaluate a word's priority score. We refer to child language literature where the difficulty of a word is modeled as a linear relation (e.g., linear regression [66, 76], linear mixed models [167]) to word-level parameters. Hence, we calculate the priority score of w_i by a linear combination of word-level parameters as follows:

$$\text{priority score } s(w_i) = k_f f(w_i) + k_c c(w_i) + k_p p(w_i) \tag{1}$$

The coefficients k_f, k_c, k_p are child-specific, as the word recognition rates are different across children even with the same word-level parameters [168]. Linear regression is performed to fit k_f, k_c, k_p such that the words that this child has heard and spoken (thus knows) (i.e., $\forall w_j \in \mathrm{S}_{\mathrm{C}} \cap \mathrm{S}_{\mathrm{NC}}$) yield higher scores $s(w_j)$ from Equation (1). In theory, the words a child knows include *receptive vocabulary* (that the child understands) and *expressive vocabulary* (that the child speaks) [30]. Currently, OSOS refers to the expressive vocabulary as those are directly observable from S_{C}.

$\boxed{E_4}$ Using the coefficients above, the scores $s(w_i)$ are calculated for the words $\forall w_i \in \mathrm{S}_{\mathrm{NC}}$ by Equation (1). Then, the words are sorted in descending order of score, yielding a prioritized list of w_i.

$\boxed{E_5}$ From the prioritized list, the Extractor takes only the words $\forall w_k \in [\mathrm{S}_{\mathrm{NC}} - \mathrm{S}_{\mathrm{C}}]$, i.e., a subset of non-child words that the child has not spoken, listed in descending order of score. We highlight that this subset, namely $\mathbf{W}_{all}$, is an ordered list compliant with the occurrence-based criteria – i.e., prioritizing the words that the child has not learned but do occur more frequently and commonly around the child. To ensure that the child has not yet acquired $\mathbf{W}_{all}$, we not only subtract S_{C} but also apply a human-in-the-loop, e.g., referring top-k words of this list to the parent, having them check off the words that the child spoke elsewhere (§6.2). For ethical precaution, a child language expert may screen inappropriate words (e.g., profanities) although existing databases may automate it [15].

$\boxed{E_6}$ From $\mathbf{W}_{all}$, the Extractor outputs top-k words; those are considered the target words for intervention tailored to this child. We further consider POS-balancing and semantic relevance in producing the target words. §6.2 elaborates on the remaining steps of target word determination.

5.3 Personalized Intervention Aid Generator (abbrev. Generator)

Our choice of intervention aids was storybooks – a common clinical practice and most children's natural routines. Our explanations below are specific to the storybook generator shown in Figure 3.

The Generator takes the target words $\mathbf{W}_i$ from the Extractor. To generate a custom story in which the target words naturally appear in the narratives, we tested multiple LLMs: GPT-3.5 & 4 [136], Bard [124], and LLaMA [173]. We settled with GPT-4 (i.e., `gpt-4-0613` snapshot) as it excelled in the naturalness of the generated stories. We elaborate on our rationale of model selection as of our development time. GPT-3.5 exhibited faster generations but inferior performances in prompt-compliance and story diversity compared to GPT-4. GPT-3.5 also tended to incorporate the target words unnaturally (e.g., using a generic word as a character's name). Bard tended to generate excerpts rather than a full story. Besides LLaMa than often underperformed GPT-3.5; we also had a concern about directly using an open-source LLM without service-level safety (which GPTs have) for children. In addition, we attempted custom story generation models [186] and plot graphs [118]. However, they showed tendencies that (1) the target words, being injected into the story keyword sequence, disrupted the overall story, and (2) the generated stories excessively comply with the given plot, producing overly stiff and disjointed writing.

Our strategies for custom story generation have been revised 3 times. Our initial attempt was to let GPT-4 generate an arbitrary story without any guidance other than the target words. It turned out to be lacking quality control and personalizability. The resulting stories varied largely in plot structure or plausibility. This method does not accommodate child-specific background, if any. Our second attempt was to provide an existing story and the target words

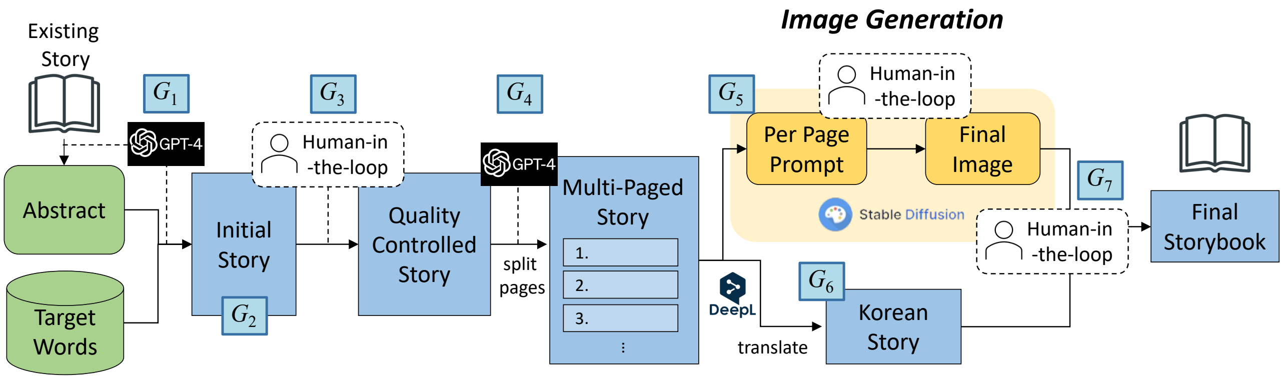


Figure 5: Detailed operation flow of Personalized Intervention Aid Generator (with implementation choice of Storybook Generator).

together, letting GPT-4 generate an altered version of the story using the target words. We discarded this method due to limited variety and unnatural incorporation of the target words. Integrating the target words directly into an existing text often resulted in similar stories or awkward episodes. One may question the copyright concerns, which we carefully discuss in §8.1.

Finally, we settled on a hybrid. Figure 5 depicts the flow of the Generator implemented in Python on our local Linux server with cloud APIs to GPT-4 [136] and DeepL [5] as well as a locally installed Stable Diffusion XL [145]. The identifiers $\boxed{G_1}$, $\boxed{G_2}$, ..., $\boxed{G_6}$ indicate the corresponding locations in Figure 5. Figure 6 showcases select major prompts that we crafted for the Generator. See the appendix for the complete set of prompts being used in OSOS, paired with examples of how it generates or changes the content.

$\boxed{G_1}$ It queries GPT-4 to extract a 3-sentence abstract from an existing story. This step removes most story-specific details, leaving only a very abstract theme. To ensure, our prompt explicitly instructs it to ignore details such as names. We empirically chose '3' sentences to balance the variety and naturalness of the story.

$\boxed{G_2}$ Now an initial story is automatically generated with the child-specific target words $\mathbf{W}_i$ and the abstract from $\boxed{G_1}$. Figure 6a lists the GPT-4 prompt for the initial story generation. This session is isolated from $\boxed{G_1}$ so that GPT-4 is unaware of the abstract's original full text. There may be many abstracts from different stories; the Generator supports multiple matching methods between $\mathbf{W}_i$ and an abstract, e.g., semantic distance [42]. The average length of a generated story was 445 words, likely affected by 'for preschoolers' instruction (Line 2). It is comparable to the children's books that we referred to. We observed that not giving an explicit length tends to produce natural stories.

$\boxed{G_3}$ An optional human-in-the-loop step may be employed, to (1) ensure the initial story incorporates the target words, and (2) is adequately structured as per *story grammar* framework [27, 109]. It is SLPs' practice to evaluate a story's structure as per story grammar, which defines the structural attributes that a story should exhibit. The attributes are: (1) Setting – time, place, characters; (2) Initiating event; (3) Goal; (4) Attempt; (5) Outcome; and (6) Reaction.

We observed that the initial story generated in $\boxed{G_2}$ is sometimes of less presence of the target words, rather simplistic, or missing a story grammar attribute. We employed human-in-the-loop iterative refinement and built an interactive web-based toolkit to expedite the iterations. Figure 7 shows the toolkit with a dozen of one-click interfaces to express a desired direction of revision, e.g., '*incorporate the target words*', '*add a conflict*', '*add more dialogues*', etc. Clicking one auto-constructs a corresponding prompt and issues it, showing the refined text interactively. Figure 8 lists the sample text snippets before and after clicking a button on the web-based toolkit, demonstrating how the generated stories are revised toward the expressed direction. Figure 6b through 6d showcase example prompts that can be issued by our web toolkit. Once issued, it takes typically 30-90 seconds to complete. See Appendix A for full prompts.

$\boxed{G_4}$ To organize the generated story into a multi-paged book, the Generator asks GPT-4 to split the text into blocks, each with 3+ sentences.

$\boxed{G_5}$ To generate illustrations to be paired with paginated texts, the Generator issues 3 image generation prompts in a chained manner as shown in Figure 6e. The first query extracts the characters from the story, which are used in the second query to generate randomized visual characteristics for each character. The third query is used to generate the base image generation prompt for each page. The responses from the last two queries are merged to create an image-generation prompt in Stable Diffusion syntax. Once the initial images are generated, a human-in-the-loop may be employed to screen possible child-inappropriate images, to steer the image styles (e.g., cartoonish), or to improve inter-page consistency (e.g., the appearance of a recurring character).

$\boxed{G_6}$ This is an extra step for non-English storybooks. Although GPT-4 supports the Korean language, its quality of native-Korean storybook generation underperformed the English counterpart. OSOS initially generates the stories in English and then translates into Korean (or another target language) using DeepL APIs [5].

$\boxed{G_7}$ The Generator compiles all paginated texts and illustrations into a single e-book PDF. For extra caution, a final human-in-the-loop step may apply here to screen the book as a whole. A minor issue (e.g., an awkward translation) is corrected inline.

The end-to-end generation of a single storybook including the human-in-the-loop steps takes about 1 hour, in which compute-only portions take 50 minutes on a server with a 2.8 GHz CPU, 256

Generate an initial story with target words

```
Remember the following words: (list of target words; typically a set of 1 noun, 1 verb, and 1 adjective or adverb)
Then use those words to write an English story for preschoolers.
Do not use the given words as names.
Do not end with a direct message to the reader. Do not make inanimate objects talk or move.
Use the summary at the end as a guideline to create the actual story.
Summary: (abstraction of a base story; typically 3 short sentences.)
```

(a) GPT-4 prompt for generating an initial story with the target words naturally embedded. Issued by our web toolkit.

Refine the story by adding conflict

```
This is the current story: (story text).
The current story is too boring and generic. Make the story more interesting.
You can make the story more interesting by adding a conflict or threat to the main character.
```

(b) GPT-4 prompt for refining the story by adding a conflict. Issued by our web toolkit.

Refine the story by adding new character

```
This is the current story: (story text).
The current story is too boring and generic. Make the story more interesting.
You can make the story more interesting by adding an interaction with a new character.
```

(c) GPT-4 prompt for refining the story by adding a new character. Issued by our web toolkit.

Increase the presence of target words

```
This is the current story: (story text).
Incorporate the following words into the story without hurting the context: (target words).
```

(d) GPT-4 prompt for increasing the presence of the target words. Issued by our web toolkit.

Generate image generation prompt for each page

```
OSOS prompt:
You will receive a story. Find all of the characters within this story.
Don't just say the name, include what they actually are, without including adjectives.
If the character is Wally, who is a rabbit, output it as Wally the rabbit.
Output them in the following format: Character 1 | Character 2 | Character 3 | .
The story is: (story text).
```

```
GPT-4 response-1:
Tom the man | Richard the man
```

```
OSOS prompt:
You will now assign the characters generated characteristics that describe what the characters look alike.
Give them characteristics that seems the most normal and fitting to the story.
If it is a human, add descriptions about their cloths.
Examples: Fox: Orange hair, Sly | Tree: Red, Tall, Maple
The story is: (story text).
```

```
GPT-4 response-2:
Tom the man: Brown hair, wears overalls, sturdy boots | Richard the man: Sharp suit, shiny shoes, slicked back hair
```

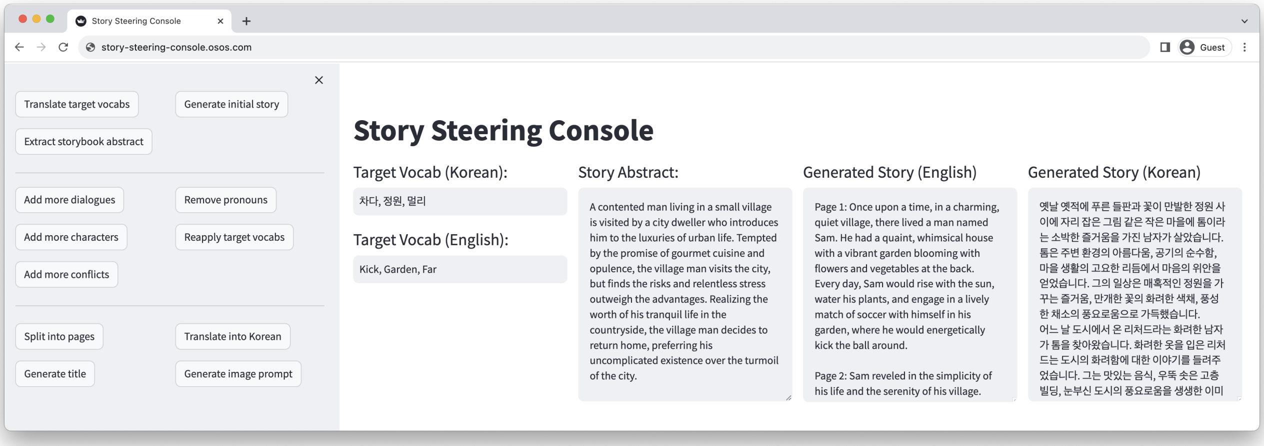
```
OSOS prompt:
You will receive a story. For each page, write me a sentence that depicts an image that can be used for the page.
The format should be [subject] [doing action], [adjective] [background subject], [scenery].
Example: [boy] [running away], [running] [dogs], [woods].
The story is: (story text).
```

```
GPT-4 response-3:
Page 1: [Tom] [tending to his colorful garden], [tranquil] [village], [green fields and lovely gardens].
Page 2: [Richard] [talking passionately to Tom], [colorful] [garden], [tranquil village].
(Output omitted)
```

```
Resulting image prompt after merging response-2 & response-3:
Page 1: man with brown hair, wears overalls, sturdy boots tending to his colorful garden, tranquil village,
green fields and lovely gardens
Page 2: man with sharp suit, shiny shoes, slicked back hair, talking passionately to man with brown hair,
wears overalls, sturdy boots, colorful garden, tranquil village
(Output omitted)
```

(e) 3 chained GPT-4 prompts to produce Stable Diffusion image-generation prompts. Issued by our web toolkit.

Figure 6: GPT-4 prompts for OSOS to generate an initial story, steer the story, and generate Stable Diffusion prompts.



Story Steering Console

Target Vocab (Korean):

차다, 정원, 멀리

Target Vocab (English):

Kick, Garden, Far

Story Abstract:

A contented man living in a small village is visited by a city dweller who introduces him to the luxuries of urban life. Tempted by the promise of gourmet cuisine and opulence, the village man visits the city, but finds the risks and relentless stress outweigh the advantages. Realizing the worth of his tranquil life in the countryside, the village man decides to return home, preferring his uncomplicated existence over the turmoil of the city.

Generated Story (English)

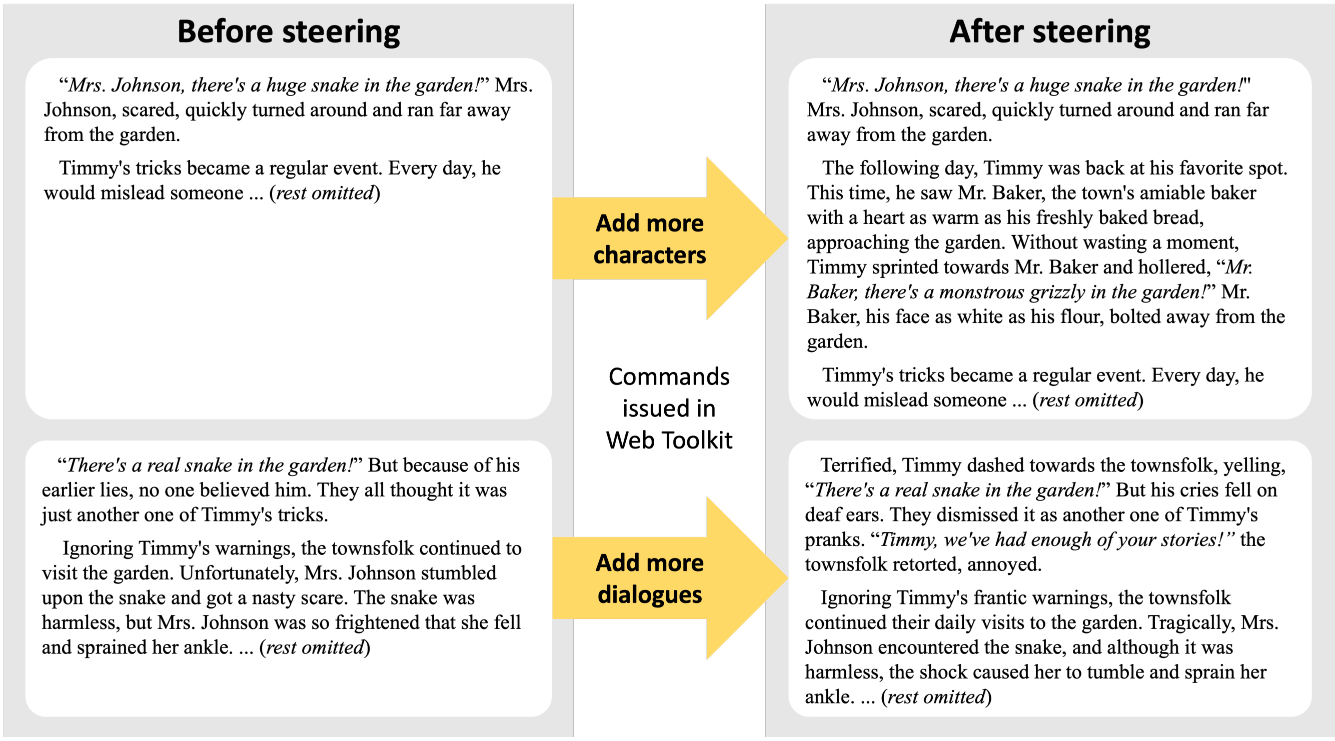
Page 1: Once upon a time, in a charming, quiet village, there lived a man named Sam. He had a quaint, whimsical house with a vibrant garden blooming with flowers and vegetables at the back. Every day, Sam would rise with the sun, water his plants, and engage in a lively match of soccer with himself in his garden, where he would energetically kick the ball around.

Page 2: Sam reveled in the simplicity of his life and the serenity of his village.

Generated Story (Korean)

옛날 옛적에 푸른 들판과 꽃이 만발한 정원 사이에 자리 잡은 그림 같은 작은 마을에 톰이라는 소박한 즐거움을 가진 남자가 살았습니다. 톰은 주변 환경의 아름다움, 공기의 순수함, 마을 생활의 고요한 리듬에서 마음의 위안을 얻었습니다. 그의 일상은 매혹적인 정원을 가꾸는 즐거움, 만개한 꽃의 화려한 색채, 풍성한 채소의 풍요로움으로 가득했습니다.
어느 날 도시에서 온 리처드라는 화려한 남자가 톰을 찾아왔습니다. 화려한 옷을 입은 리처드는 도시의 화려함에 대한 이야기를 들려주었습니다. 그는 맛있는 음식, 우뚝 솟은 고층 빌딩, 눈부신 도시의 풍요로움을 생생한 이미

Figure 7: Interactive web-based toolkit for story steering.



Before steering

"*Mrs. Johnson, there's a huge snake in the garden!*" Mrs. Johnson, scared, quickly turned around and ran far away from the garden.

Timmy's tricks became a regular event. Every day, he would mislead someone ... (*rest omitted*)

"*There's a real snake in the garden!*" But because of his earlier lies, no one believed him. They all thought it was just another one of Timmy's tricks.

Ignoring Timmy's warnings, the townsfolk continued to visit the garden. Unfortunately, Mrs. Johnson stumbled upon the snake and got a nasty scare. The snake was harmless, but Mrs. Johnson was so frightened that she fell and sprained her ankle. ... (*rest omitted*)

After steering

"*Mrs. Johnson, there's a huge snake in the garden!*" Mrs. Johnson, scared, quickly turned around and ran far away from the garden.

The following day, Timmy was back at his favorite spot. This time, he saw Mr. Baker, the town's amiable baker with a heart as warm as his freshly baked bread, approaching the garden. Without wasting a moment, Timmy sprinted towards Mr. Baker and hollered, "*Mr. Baker, there's a monstrous grizzly in the garden!*" Mr. Baker, his face as white as his flour, bolted away from the garden.

Timmy's tricks became a regular event. Every day, he would mislead someone ... (*rest omitted*)

Terrified, Timmy dashed towards the townsfolk, yelling, "*There's a real snake in the garden!*" But his cries fell on deaf ears. They dismissed it as another one of Timmy's pranks. "*Timmy, we've had enough of your stories!*" the townsfolk retorted, annoyed.

Ignoring Timmy's frantic warnings, the townsfolk continued their daily visits to the garden. Tragically, Mrs. Johnson encountered the snake, and although it was harmless, the shock caused her to tumble and sprain her ankle. ... (*rest omitted*)

Figure 8: Sample story snippets demonstrating how a command on the web-based toolkit revises the generated stories.

GB memory, and an RTX 3090 GPU. To break down the 50 minutes of computing time, generating one storybook's worth of text is done by a single prompt, taking less than 5 minutes. However, the images are generated in batches, taking (4.5 minutes per 20-image batch) × (10 pages per book on average) = 45 minutes in total. As a part of human-in-the-loop steps, the best image out of each 20-image batch was chosen per page. It is worth the time as images are essential for children's books in attracting attention and conveying the context [133]. Our results are also supportive of the images' contribution to the children's interest in the books (§7).

6 IN-THE-WILD DEPLOYMENT TO FAMILIES

To evaluate, we deployed OSOS in-the-wild to actual homes. §6.1 describes the participants and the overall study structure. §6.2 and

§6.3 describe the deployment setup of Stage 1 and 2, respectively. §7 reports the results and findings.

6.1 Participants and Study Structure

We recruited 10 families from local kindergartens and language clinics. All families were native Korean-speaking. In each family, one child and one parent participated. The children were 6 females and 4 males; their min/mean/max ages were 50/62.2/71 months (i.e., 4-5 years old), respectively. C1, ..., C10 denote each child, and P1, ..., P10 denote each respective parent. We targeted this age group in line with the emphasis on vocabulary assessment for preschool ages [57], their routinization of book-reading with parents, and pronunciation development reasonably recognizable by STT APIs. Prior to joining the study, each child was referred to an SLP and evaluated through Receptive and Expressive Vocabulary Test (REVT) [103] – a Korean extension of Expressive Vocabulary Test [184] to ensure that the child meets our participation criteria (i.e., children with typical development). To clarify, this REVT is not intended for pre-/post-test to evaluate the efficacy of OSOS deployment. Furthermore, it is advised *not* to administer standardized tests repeatedly within a certain interval (e.g., 6 months) as the learning effect would result in over-estimation [87].

The total deployment period was 4 weeks, consisting of 2 stages:

1. Voice Collection (earlier 2 weeks): OSOS collects the child-specific language environment at home.
2. Parent-Child Storybook Reading (later 2 weeks): Parent and child read OSOS-generated personalized storybooks.

After concluding both stages, each parent was invited to a post-deployment questionnaire and a 1.5-hour individual exit interview. Each family was compensated a USD 200-worth amount.

The whole study plan was approved by our university IRB (PIRB-2023-020, PIRB-2023-023). We took special precautions in study designs. While OSOS aims to help children's vocabulary learning, we implemented multiple safety layers to prevent the slim possibility of embarrassing deviations of the generative AIs. SLPs checked new storybooks before delivery to the families. The parents checked newly delivered books before reading with the child. Regarding the privacy concerns in voice profiling, we ensured the parents had full control over when to record or not, and obtained their consent. §8.3 discusses further privacy optimizations for productization.

6.2 Deployment Setup: Stage 1. Voice Collection

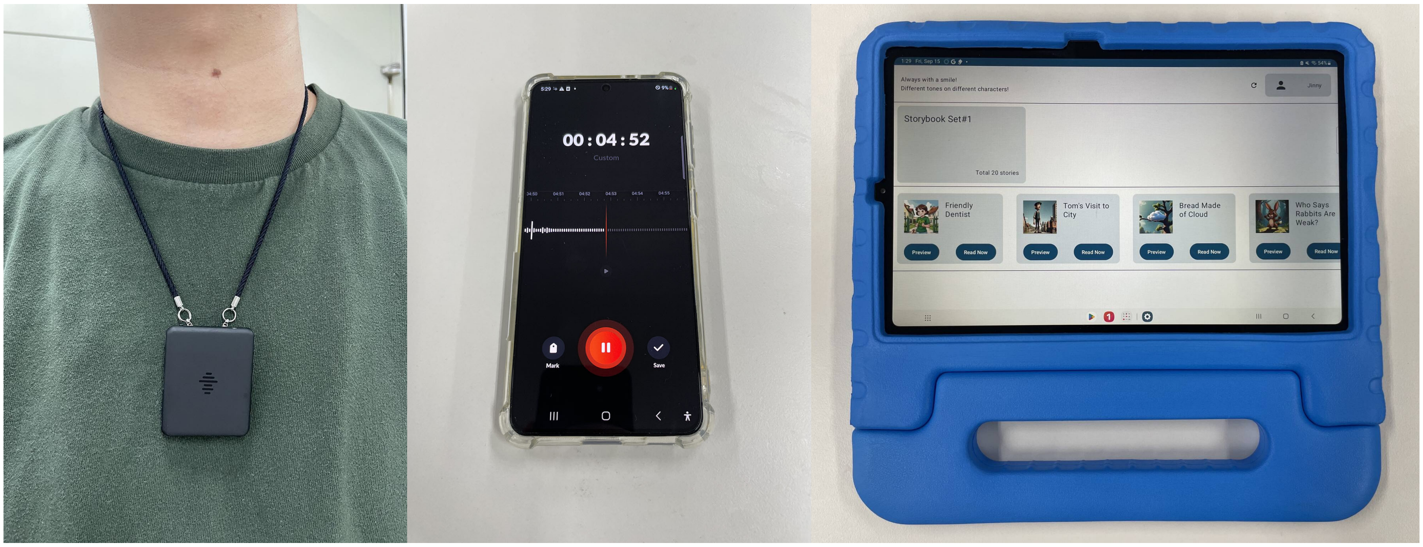
To each family, we deployed a pair of voice recorders – one wearable and one stationary (Figure 9). The parents were asked to record their daily life with the child for about (10 hours a week) × (2 weeks). They freely chose when to begin and stop recording, but were asked to refrain from recording when the conversation was inactive.

During the same 2 weeks, we also asked the parents to engage in tablet-based storybook reading. Each child-parent dyad selected favorite books that they own, which were scanned and uploaded to the tablet we provided. An average of 28 e-books are provided to each family. We asked them to have about 20 minutes of tablet-reading sessions per day. This activity was to mitigate possible novelty effects in Stage 2 – when reading OSOS-generated books using the tablet.

All the voice recordings and tablet-readings were in-person activities. All the devices provided to each family were of identical models. To validate the reliability of auto-transcription, we sampled 30 minutes of recordings from each family. Comparisons to an SLP's manual transcriptions show a 97% precision of 77% recall, implying that our STT APIs operate in a highly conservative setting. With such high precision, we can be almost certain that the vocabulary in the transcription did occur in their home. Meanwhile, we find that modest recall mainly comes from faint YouTube or TV in the background. As for our POS analyzer, the literature reports an F1-measure of 0.83 for a standard Korean corpus of 1.3 million sentences [39]. Empirically, however, we observed rather stable POS tagging; we conjecture the reasons that (1) the reported F1-measure is for every morpheme, but many of the morphemes are not the intervention target of OSOS and thus filtered out; and (2) the standard corpus includes neologism which rarely appears in children's books. Note that the Korean language, due to its agglutinative-dominant properties [164], poses inherent difficulties in POS analysis. For English which is much less agglutinative and largely isolating [45, 69], its POS analyzer attains over 97% of accuracy [123], implying higher performance of OSOS in English environments.

Upon concluding the 2 weeks of voice collection, their recordings were processed by the Profiler (§5.1) and Extractor (§5.2). From each family, the Extractor returned a prioritized list of words that frequently and commonly occurred in non-child speech but did not appear in the child's speech (detailed in §5.2). These words are considered the priority vocabulary for intervention as per the occurrence-based criteria [111]. As explained in §5.2, comparisons between different criteria options are not a goal of this paper; the occurrence-based criteria are chosen as they are widely exercised in clinical practices. We placed a human-in-the-loop here to remove possible (although rare) child-inappropriate words from the list.

Figure 10 depicts how the target words for each child are selected and divided. We took high-ranked words from the prioritized list, and referred those words to the parent to check off any words (as they recall) that the child has spoken elsewhere (but not captured in the recordings). This parental survey is a valid clinical procedure in standard tools such as M-B CDI [57], as explained in §2.2. The remaining words after parent filtering are divided by POS, forming priority-ordered per-POS lists of words that the child does not know. To understand the efficacy of OSOS compared to the baseline (where the child naturally acquires words occurring in their environment), we need two groups of words of equivalent conditions – (1) the *target words* that personalized storybooks will be generated with, and (2) a control group of words, namely *control words*, whose priorities are equivalent with the target words but are not fed to the storybooks. Hence, we sample target words and control words alternatively in descending order of priority. Specifically, the target words are taken from the 1st, 3rd, 5th, and 7th-ranked positions from each POS group of nouns, verbs, and adjectives/adverbs. The control words are taken from the even-numbered ranks from each POS group. In total, we obtained 12 target words and 12 control words per child. These 12 target words are further divided into 4 subsets of T_1, T_2, T_3, T_4 each with {1 noun, 1 verb, 1 adjective/adverb} based on semantic distances therein. Obviously, these target words are personally prioritized for each child; the target words from different children are unrelated and almost completely different



(a) Wearable recorder

(b) Stationary recorder

(c) Custom reading app on an Android tablet

Figure 9: Two types of recorders and a tablet deployed.

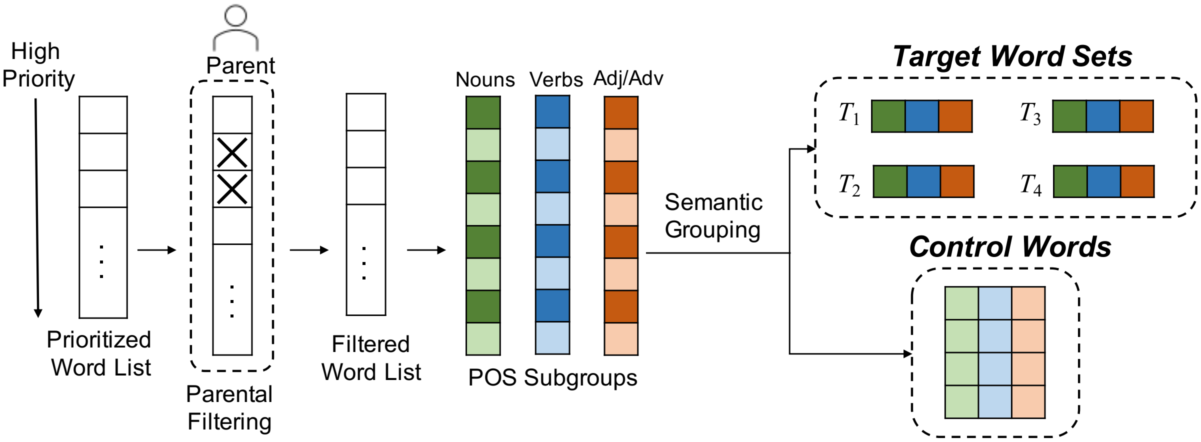


Figure 10: Words selection setup: selection and division process of target words and control words.

(i.e., 101 unique words in the union of $12 \times 9 = 108$ target words for all 9 children). **We did not disclose the target words nor the control words to the parents** until completion of deployment.

6.3 Deployment Setup: Stage 2. Parent-Child Storybook Reading

The family of C5 dropped out due to too little amount of voice recorded. A total of 9 families proceeded to Stage 2.

Figure 11 depicts our storybook generation setup. As described in §5.3, for each family, we distilled 20 abstracts (3 sentences each) from 20 different storybooks they owned. These 20 abstracts are matched with 4 personalized target word sets (T_1, T_2, T_3, T_4) obtained from §6.2, resulting in (5 books for each set of 3 POS-balanced target words) × (4 sets) = (total 20 books per child). For optimal matches, stable matching algorithm [151] resulted in the 5 books sharing the same 3 target words to be sorted in terms of the semantic distance between an abstract and a target word set.

As a result, OSOS generated a gross total of 180 storybooks for 9 children (20 books per each). For each child, each target word is embodied through 5 different storybooks, where each target word appeared an average of 1.24 times per book. Before distribution, all storybooks were screened by an SLP to determine their child-appropriateness.

Figure 12 demonstrates sample pages excerpted from two OSOS-generated storybooks. The sample books were generated with the target words {**garden, kick, far**}. The book shown in Figure 12a and 12b was generated from the abstract of '*The Boy Who Cried Wolf.*' The book in Figure 12c and 12d was generated from the abstract of '*The Town Mouse and the Country Mouse.*' Both original stories belong to '*Aesop Fables*' in the public domain.

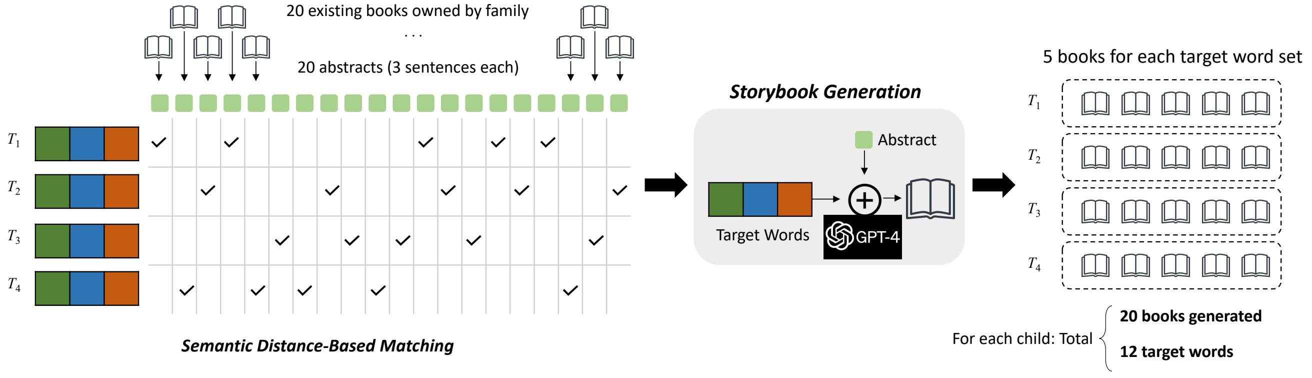
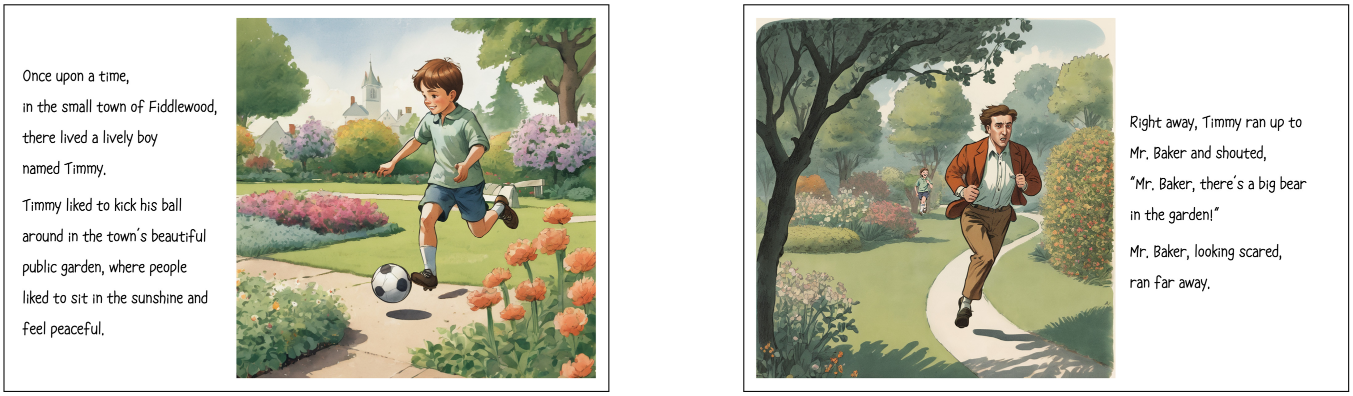


Figure 11: Storybook generation setup: combining abstracts and target word sets, generating 5 storybooks for each word set



Once upon a time, in the small town of Fiddlewood, there lived a lively boy named Timmy.

Timmy liked to kick his ball around in the town's beautiful public garden, where people liked to sit in the sunshine and feel peaceful.

Right away, Timmy ran up to Mr. Baker and shouted, "Mr. Baker, there's a big bear in the garden!"

Mr. Baker, looking scared, ran far away.

(a) Page #1 from a generated book auto-titled by GPT-4 '*When Fiddlewood Stopped Listening*'

(b) Page #5 from a generated book auto-titled by GPT-4 '*When Fiddlewood Stopped Listening*'

Once upon a time, in a quiet village far, far away, there lived a cat named Sam.

He had a cozy house with a bright garden full of flowers and veggies at the back.

Every day, Sam would wake up with the sun, water his plants, and have fun kicking his soccer ball around in his garden.

Sam stood on a rooftop, gazing at the village far beyond the city's limits.

Sam missed his garden, his daily soccer practices, and the gentle sounds of birds singing.

The city had lots to see, but it didn't feel as homey as his cottage.

(c) Page #1 from a generated book auto-titled by GPT-4 '*Village Cat Visits the City*'

(d) Page #8 from a generated book auto-titled by GPT-4 '*Village Cat Visits the City*'

Figure 12: Sample pages taken from OSOS-generated storybooks (not necessarily consecutive). All books are generated to incorporate the target word set of {`garden, kick, far`} as per our {noun, verb, adjective/adverb} configuration. The book with pages (a) and (b) was generated from the abstract of '*The Boy Who Cried Wolf*.' The book with pages (c) and (d) was generated from the abstract of '*The Town Mouse and the Country Mouse*.' Note that both original stories belong to '*Aesop Fables*' which are in the public domain. The illustrations are also OSOS-generated with a few high-level human-given options such as '`storybook illustration style`' (for (a) and (b)) and '`fantasy art style`' (for (c) and (d)).

To distribute storybooks and collect statistics on their reading activities, we built a custom Android reading app that runs on the tablet (Figure 9c). This app is capable of updating the storybooks delivered online, logging child-parent dyads' reading patterns, e.g., per-page retention, how many times each book was read, voice-recording during active reading sessions, etc.

We asked each family to (1) read every personalized book at least once through Stage 2, (2) have at least 6 reading sessions a week, (3) spend about 20 minutes per reading session, and (4) read at least two different books per reading session. As mentioned in §6.2, they read the books without knowing the target words. We chose not to *explicitly* ask the parents to exercise *read-aloud* [52, 128], because: (1) We observed from Stage 1 that read-aloud is already their routine. (2) To facilitate OSOS naturally adopted to their routines, we did not enforce a specific reading style; we simply asked them to read together as they usually do. Later, upon analyzing the voice recording during reading sessions, we observed many read-aloud activities and interactions, e.g., the parent reads aloud to the child, they ask each other about the book's content, the parent relates a book's episode to the child's daily life, etc.

7 RESULTS AND FINDINGS

From the deployment, we collected the following data from each family, listed in a chronological order of collection:

1. Voice recordings at home – 28.7 hours on average. (collected during Stage 1)
2. Reading activity logs and recordings from our custom tablet app (collected during Stage 2)
3. 1st parental survey on the profiled words, where the parent checked off the words that the child spoke elsewhere (at the end of Stage 1; detailed in §6.2)
4. 2nd parental survey on the profiled words, where the parent checked off the words that the child spoke elsewhere – repeated the same survey as in (3). (at the end of Stage 2)
5. Post-deployment questionnaire (at the end of Stage 2)
6. 1.5 hours of individual semi-structured interview with each parent (at the end of Stage 2)

7.1 Voice and Reading Statistics

Stage 1: Daily-life verbal activities. Analyses of the voice recordings (Stage 1) and reading logs from our custom tablet app (Stage 2) reveal the following statistics. All values are family-average unless otherwise stated. From Stage 1, we collected 28.7 hours of voice per family, in which child speech accounts for 26.8%. We found $N_{total} = 2276$ unique words in total. $N_{\mathrm{C}} = 705$ unique words appeared in child speech, while $N_{\mathrm{NC}} = 1973$ unique words in non-child speech. At the end of Stage 1, the parent was referred to an average of 567 words – the prioritized words that OSOS extracted from their real-life voice activity. The parent filtered out an average of 369 words (i.e., the child knows).

Stage 2: Reading activities of generated books. In Stage 2, each family spent 15.3 minutes per day reading OSOS-generated storybooks. Most books were read more than once on different days (2.2 days on average; most read: 6 days). They spent an average of 9.2 minutes on each book. We further analyzed the voice recordings from our reading app. Except for two children (C4 and C9), the children's utterances accounted for 4.7% of total utterance time, mainly for occasional questions and interaction with the parents while the books were being read-aloud by the parents. C4 and C9 exhibited outlying utterance times (83.2% and 19.8%, respectively) as they had higher literacy levels; C4 read-aloud most books herself while occasionally conversing with the parent, and C9 read-aloud together with the parent. The children except C4 and C9 made an utterance once every 39 seconds (i.e., roughly once a page) where each utterance was 3.42-sec-long on average. Major types of child-utterances include: (1) Asking a question about an unknown word in the book, e.g., C1: "*What is 'shivering'?*", P7: "*Do you know what 'merchant' means? Someone selling things.*" (2) Relating a book's content with the child's real-life episode, e.g., P8: "*Do you feel it difficult to finish something, like [the book character]?*" (3) Talking about an image in the book, e.g., P8: "*Where would [book character] hide the car? Where would you hide it?*" (4) Reflection after finishing a book, e.g., P1: "*What was the best part?*", C1: "*The wedding.*"

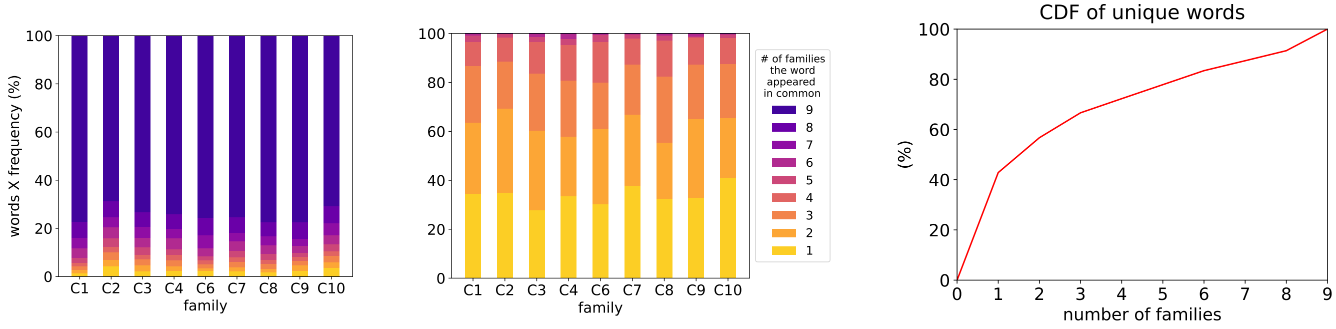
7.2 Vocabulary from Daily Life and into Generated Books

Difference of language environment across families. To verify how different (or similar) the language environments are across families, we analyzed the words profiled in Stage 1. Figure 13a depicts the occurrence distribution (i.e., words × frequency) in terms of commonality across families, indicating clear frequency-dominance of the common words occurring in all 9 families. However, those common words constitute only 8.6% of the total unique words; Figure 13c depicts the CDF of unique words, suggesting highly skewed occurrences of a few frequently spoken words. Indeed, the common words largely consisted of stopwords, function words, and infant-level words (e.g., do, is, mom, eat, etc).

Now we analyze the distribution of the words that each child has *not* spoken – which are deemed that they need to learn, as explained in §5.2 (denoted $[\mathrm{S_{NC}} - \mathrm{S_C}]$). As a result, we observed $N_{\mathrm{NC-C}} = 1779$ unique words out of the total $N_{total} = 2276$ words, implying that the child presumably may not know up to 78.2% of the vocabulary that has appeared in that child's home. Figure 13b depicts the occurrence distribution of such words; it sharply contrasts with Figure 13a. 43% of the total occurrences are from unique words that appeared only in one family; the words that appeared in 5 or more families constitute only 28% of the occurrences. This diversity may seem striking given that our participating families were all native Korean-speakers. But it would be understandable as the vocabulary spoken in a family is a living set, closely reflecting the recent activities of family members (e.g., say a family has returned from hiking; their vocabulary will naturally reflect many nature- or plant-related words).

Overall, these findings are supportive of our motivation that children's daily language environment would largely differ across families, necessitating personalized assessment and intervention. We admit that this is a limited 20-hour snapshot per family. Still, extended profiling may bring both effects – previously family-unique words may appear in other families, but previously unseen words may also appear. Thus, we speculate that the motivation of OSOS would still hold with extended profiling.

Reflection of personalized target words onto generated storybooks. We observe that each target word textually appeared at a relatively similar frequency in their corresponding books, i.e., 1.24 times per book on average (std: 0.35). Given our deployment setup that incorporated each target word set into 5 different storybooks



(a) For the entire words that appeared in each family (i.e., $S_{NC} \cup S_C$)

(b) For the words that the child has not spoken yet (i.e., $S_{NC} - S_C$)

(c) Cumulative distribution of unique words along the number of families where the word commonly appeared.

Figure 13: Word × frequency distribution along with the number of families where the word commonly appeared, and Cumulative distribution of unique words along the number of families where the word commonly appeared.

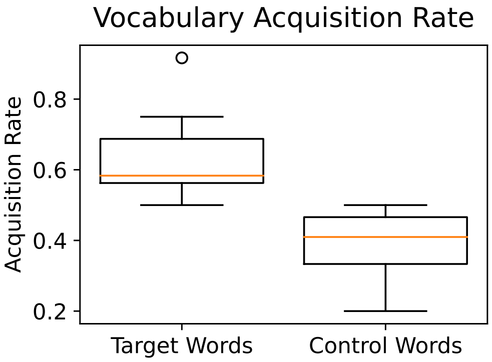


Figure 14: Vocabulary Acquisition Rate

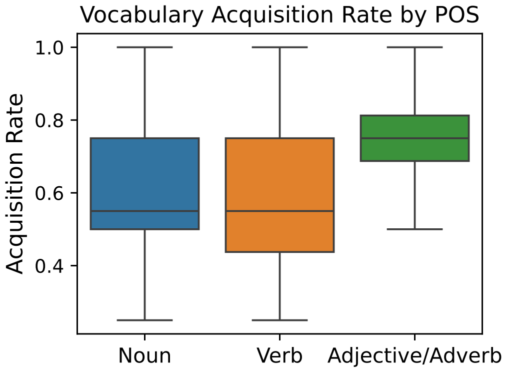


Figure 15: Vocabulary Acquisition Rate by POS

and our reading app logs of how many times each book was read again, each target word would have been exposed to the respective child at least 13.64 times. On the contrary, each target word's visual reflections in the corresponding books' images were fewer and more fluctuating, i.e., 0.65 times per book (std: 1.68). We speculate on two reasons. First, the target words could be of a type seldom visualized. On a page saying "*It is my* <u>*honor*</u> *to meet you.*", the target word "*honor*" is unlikely visualized. For "*Sam traveled around the* <u>*world*</u>*.*", the generated images tend to show traveling in one place, not all over the globe. Second, we observed that the image models tend to abide by certain ways of visualizing the prompt while lacking the flexibility of nuanced expression. For "*Ellie is* <u>*leaving*</u> *the village by carriage.*", our Stable Diffusion model always generated their frontal view, while a human illustrator may express 'leaving' by drawing their back. Still, it does not undermine the importance of images. §7.4 reports that the images are highly correlated with the children's interest in the books, which is in turn correlated with their acquisition of target words.

7.3 Vocabulary Acquisition by Children

Vocabulary acquisition rates: target words vs. control words. Upon concluding Stage 2, each parent was given a vocabulary list that was identical to what they checked off at the end of Stage 1. This list includes the target words and control words not known to the parents. Once again, they filtered out any words that they thought their child had spoken so far. Then, we cross-referenced the parent's checklist and the words in our target and control sets, respectively. Figure 14 shows the results that, after Stage 2, each child has acquired an average of 64% of the target words while acquiring 39% of the control words. As OSOS did not deliver the words in the control sets, they would have been acquired elsewhere. Note that the target words and control words were equivalently top-ranked in terms of their real-life occurrences. Mann-Whitney U Test confirms the difference in the acquisition rates ($p = 0.0017$). We believe that applying different word-selection criteria (e.g., prioritizing infrequent words [25, 70, 132]) may enlarge the target-control difference of acquisition rates.

Vocabulary acquisition rates: difference by POS. We examined whether the acquisition rate of target words differed by part-of-speech category: nouns, verbs, and adjectives/adverbs. Figure 15 shows the result. While we found that target words of adjective/adverb had slightly higher acquisition rates than those in other POS, the differences were not statistically significant. These observations somewhat deviate from existing literature that specific POS categories are easier to acquire than other categories – e.g., nouns are acquired earlier than verbs [24, 58, 64] and even adults perform well in naming tests producing nouns compared to those producing verbs [29]. It calls for further investigation on the efficacy of the generated intervention with non-noun target words.

Vocabulary acquisition rates: semantic distance factors. We further analyzed the possible difference in acquisition rates with respect to: (1) the semantic distance between the target word set and its matched abstract that resulted in a storybook; (2) the semantic distance between that abstract and the semantic centroid of the existing storybooks each family owned. Counterintuitively, either case did not show significant correlations. It could be explained that the children are not necessarily attracted to the storybooks they were familiar with (partly supported by the parents' interviews). Rather, novel custom storybooks that deviate from a certain norm may be compelling to them.

Vocabulary acquisition: difference by word attributes. To examine the factors influencing word acquisition, we analyze the impact of various word attributes. For each target word, we categorize it according to (1) the role of the character which the word is used with, (2) its presence in dialogues, and (3) the concept (concrete or abstract). Figure 16 illustrates the distribution of acquired and non-acquired words across these categories. We did not find significant differences in the distribution of character roles and dialogue presence between acquired and non-acquired words. However, Figure 16c shows that more than 60% among the acquired vocabularies were concrete words, while 42.9% of the non-acquired vocabularies were concrete. This observation is consistent with the widely accepted understanding that concrete words are easier to acquire than abstract words [158].

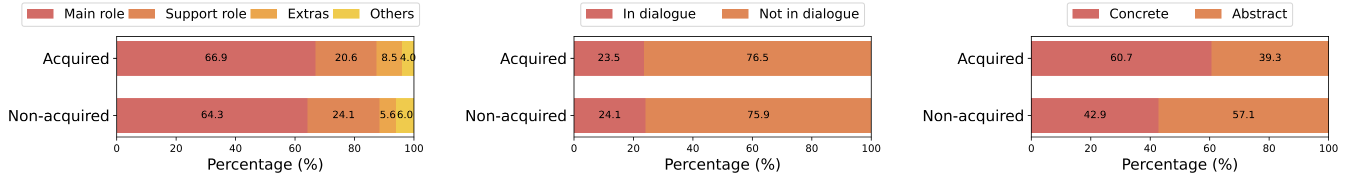
Vocabulary acquisition rates: individual differences. When comparing the difference between each child's acquisition rate of target words and control words, the acquisition rate of target words exceeded that of control words by a range of 0 to 42 percentage points (median: 24% points). This variance did not exhibit a significant correlation with either the time each family spent reading OSOS's storybooks during the deployment or the proportion of child speech during storybook reading sessions. We believe these are attributable to the child's intrinsic factors in speech-language pathology – including internal factors such as working memory [20, 53, 63], phonological processing [178], attention [120], as well as environmental factors like socioeconomic status [60, 77], parenting behaviors [121]. It is known that these contribute to individual differences in vocabulary acquisition, and discrepancies in acquisition rates between children are likely to emerge even with controlled interventions done by speech-language pathologists. Furthermore, we have not controlled the difficulty of the word, so the factors like pronounceability of each word may impact its ease of acquisition. Future extensions of OSOS would consider these factors to improve personalization performance.

7.4 Post-deployment Parental Questionnaire

Parent responses on the target words. Now we disclosed to the parents what the target words were. They were given a questionnaire with 5-pt Likert scale questions about each target word (and thus appeared in their generated storybooks), such as "*Have your child been exposed to this word often?*", "*Would your child need this word soon in daily life?*", "*Does this word appear in your existing storybooks often?*", and so on. (5: Strongly Agree) This questionnaire is adopted from the literature on educational benefit-and-cost in vocabulary selection [21]. Figure 17 depicts their responses for select questions. Figure 17a through 17c show overall high agreements, supporting that they appreciate the presence, appropriateness, and necessity of the target word, with respect to their own child's daily living context. Interestingly, Figure 17d shows overall neutral responses regarding the occurrences of that word in their existing storybooks. We elicit two major implications. Firstly, our personally generated storybooks stand out in that they deliver vocabulary less likely seen in ready-made storybooks. Secondly, the parents acknowledge that the personalized target words reflect well their children's daily lives and near-future needs, which have been less covered by existing books. Along with the results of the Post-Study Questionnaire, we referred to standard vocabulary lists to see if the words targeted for learning in the OSOS reflect individual differences not covered by standards. We first cross-checked our target words with the standard word list used for assessment. We examined Receptive and Expressive Vocabulary Test (REVT) [103], a standard measure of receptive and expressive vocabulary in Korean. Then we also consulted a list of high-frequency substantive [35] and predicates [135] known by average children aged 2 to 5 years. None of the target words were found in the REVT, and only about 5% were present in the high-frequency word list. These findings also support that the resultant OSOS system is in line with the original intent - capturing everyday words beyond the standard into the scope of assessment and intervention.

Child's interest in AI-generated books: contents, images, and vocabulary acquisition. The questionnaire also asked the parent to rate their child's overall interest shown in each of the 20 books, as well as specific interest in each book's textual content and pictorial content, respectively (5-pt Likert scale). Spearman's rank correlation analysis indicates that a child's text-specific and image-specific interests are both highly correlated to the child's overall interest in that book ($\rho = 0.961$ and 0.693, respectively; $p < 0.001$ in all cases). Furthermore, we found that a child's interest level in a book is positively correlated to the child's acquisition rate of the target words in that book ($\rho = 0.418; p = 0.011$). These observations support that images are indeed important in children's books – being correlated to the child's interest in the book, and in turn, to the acquisition of the target words that the book was generated with. Therefore, we believe images are worth a rather long generation time, as discussed in §5.3.

Child's interest in AI-generated books: personalized factors We found individual children's diversity in correlations between various story-related factors and their interests in the book using point-biserial correlation analysis. For example, C10 exhibited a higher interest in books whose protagonist is a similar-aged child ($r_{pb} = 0.50, p = 0.025$) while C2 and C3 liked books with an adult- ($r_{pb} = 0.53, p = 0.016$) and an animal-protagonist ($r_{pb} = 0.59, p < 0.01$), respectively; C8 liked a book without an explicit lesson ($r_{pb} = 0.49, p = 0.029$). Referring to the earlier finding that a child's level of interest in a book is correlated to their

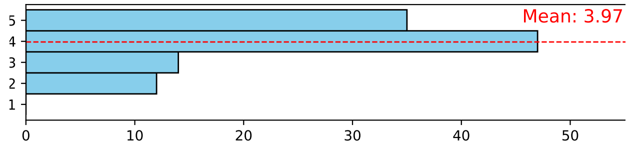


(a) Distribution comparison of embedded words related to character roles in acquired and non-acquired words.

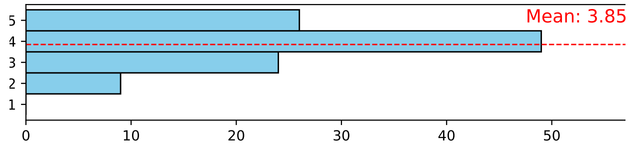
(b) Distribution of embedded word used in Dialogue vs. Non-Dialogue context in acquired and non-acquired words.

(c) Distribution of concrete and abstract words in acquired and non-acquired words.

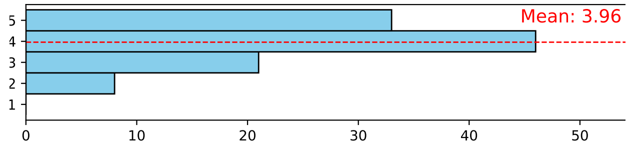
Figure 16: Distribution comparison of acquired and non-acquired words across different vocabulary attributes.



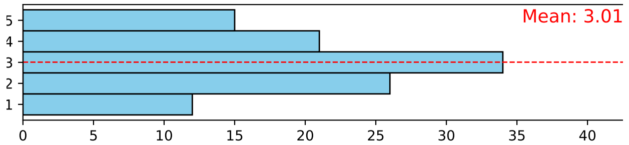
(a) Q: *Have your child been exposed to this word often?*



(b) Q: *Is this word appropriate to your child's language level?*



(c) Q: *Would your child need this word soon in daily life?*



(d) Q: *Does this word appear in your existing storybooks often?*

Figure 17: Parents' responses to select questions in the post-deployment questionnaire. All questions are on a 5-point Likert scale (1: Strongly disagree, 5: Strongly agree). The vertical red lines indicate the mean lines.

acquisition of the target words in that book, these findings shed light on another layer of personalization to maximize the targeted vocabulary acquisition – i.e., crafting the story-/image-generation prompts to include or exclude child-specific factors that are positively or negatively correlated to one's interest, respectively. As the dimensionality of personalized factors would grow, constrained generation techniques for LLMs [187, 196] would be helpful to ensure adherence to the complex constraints.

7.5 Exit Interview with Parents

As the last step, we conducted a 1.5-hour-long individual semi-structured exit interview with each parent. The interviews are analyzed by 2 researchers. Below, we report the high-level themes [170] and the representative quotes.

Learning vocabulary with tailored stories. Most parents said that the target vocabulary was naturally blended into the story, while a few pointed out slightly awkward expressions. P9: "*I haven't seen words like 'adjustment' or 'evaluation' in existing books. [The child] hasn't spoken these words, but I believe he has heard them. He seemed to smoothly understand those words in line with the context.*" P1: "*So, the word 'timing' popped up (from a book). I would get a brain freeze if she asked what it is out of thin air. But it was much easier as it came with the context.*" P2: "*I felt most words went well along with the story. But a few times I saw some words used in a slightly awkward way.*"

Responses after being told the target words. Recall that we kept the target words from the parents until the post-deployment questionnaire. They responded that there were exotic words unlikely to appear in existing storybooks. Furthermore, the parents noticed their children often saying novel words that they had not. P10: "*(Upon seeing the target words for the first time) Oh, I see 'sticky', 'worry', and 'order'. Now I recall she's been saying those words recently. Perhaps she picked them up from the books.*"

Limitations in AI-generated stories. The parents commented that their children did find interest in the AI-generated books as those were brand-new stories, but there were limitations. P9: "*Many books start with similar expressions, 'once upon a time ...', 'there was a girl ...'*." We observe this is a typical bias of GPT perhaps due to a large training corpus of existing stories. P3: "*(The generated storybooks) tend to explain too much, leaving few things the child finds curious about.*" We believe this is partly because we built OSOS to abide by the story grammar, which emphasizes causality. In addition, our design intent of OSOS favoring smaller human effort in human-AI co-creation process, coupled with yet-to-be perfect generative AI, may be partly responsible. We believe that advancements in AI will gradually mitigate this issue. In the interim, we can resort to adding human involvement, such as augmenting existing content for fine-grained steering or revision [6, 9, 13, 163].

Limitations in AI-generated illustrations. Some parents pointed out the single-subjected tendency of AI-generated images. P9: "*(Existing) books show many things altogether, even off-topic stuff. That triggers many conversations. (Generated) pictures lack such richness.* " P3: "*[A character] didn't look the same age over pages. Still identifiable, though.*" We note that identity consistency is ongoing research in generative models [67, 148].

Language environment profiling beyond vocabulary. The parents called for richer reflection of the surrounding context, beyond isolated vocabulary. P2: "*We went hiking and talked about a landslide, but (the child) didn't understand. I'd like OSOS to pick up the context and create a story.*"

Alternative criteria of target vocabulary selection. The parents suggested various criteria for target word selection, e.g., semantically relevant words to what the children already know, or words that the child would hardly encounter in daily life. P4: "*We always goes to malls; she doesn't know markets.*" P3: "*Now that he knows 'red', I'd like books that feature words like 'crimson', 'scarlet', 'maroon'* ..." Note that OSOS can flexibly switch the target word selection criteria. In particular, the parents' suggestions are in line with the alternative criteria we referred to in §5.2 [25, 70, 132].

8 DISCUSSION

In this section, we discuss various implications, limitations, and future work.

8.1 Issues and Considerations about Copyright

Taking the full text of an existing story will raise copyright issues unless exercised with caution. We acknowledge that copyright laws often allow exemptions for noncommercial scientific or educational use. We also note that academic communities have accepted publications with AI-generated works presenting the style of commercial studios [37, 119]. Still, further extensions beyond a research prototype will be limited. As shown in §5.2, OSOS refers to only a 3-sentence abstract, where little details of the original story remain. OSOS does not see the original illustrations. This may mitigate the concerns but not be an ultimate resolution, as copyright infringement would be determined by similarity or intent.

Our resolution is to limit OSOS to work with one's owned stories for personal use, which is an exemption in many countries. We can also use the stories in the public domain [2]. Every year, many books enter the public domain as their copyrights expire[1]; the U.S. Library of Congress lists books that are *free to use and reuse* [8].

If one extracts an abstract from a copyrighted story for commercial use, they should consult a legal professional. Nevertheless, OSOS does not intend copyright infringement. Our purpose remains noncommercial, i.e., to explore a new interdisciplinary scientific opportunity between computing systems and language pathology.

8.2 Deployment Population: Limitations and Future Extensions

Current deployment target. Our deployment did not mainly target children with language delay. This was to refrain from making a full-scale deployment of an experimental system onto a clinically sensitive group. By nature, the beneficiaries of OSOS include every child regardless of delay. Based on the lessons from this deployment, a future extension will include fine-tuning the word selection criteria, new dimensions diversifying storybooks, and deployment to children with vocabulary delay under SLP supervision.

OSOS for mitigating the '*30-million-word gap*'. We envision that the ability of OSOS – systematically reflecting a child's language environment and extracting the words they would need most therein – would be promising for narrowing the *'word gap'* [74]. It is known that a child in a family with high socioeconomic status (SES) hears 2,153 words per hour, while a child in a low-SES family hears only 616 words, making *30-million-word gap* in 4 years [73]. By switching into alternative target word criteria [25, 70, 132] that seeks outside of the child's own environment, OSOS would promisingly contribute towards equalizing the word sets that children would likely hear despite different SES. Furthermore, a similar benefit may be applicable to the children in immigrant families [108].

Study designs that indirectly involve the participating children. As per clinical protocols, either direct or indirect assessment methods can be exercised to evaluate a child's vocabulary learning [143]. Direct methods show picture cards to the child and elicit their description of the picture [51], while indirect methods ask the parent whether the child knows certain words [57]. In designing our deployment, the SLPs advised that indirect assessment is more appropriate. The reasons include: (1) Direct assessments should be spaced sparsely enough (e.g., 6+ months) [87] because repeated assessments using the same tool likely develop practice effects. (2) The assessment itself will ask the target words and control words, causing unintended learning. (3) Direct assessment by the researchers unfamiliar to the child may result in underestimation. For these reasons, the SLPs recommended the indirect method via parents, which is a well-established protocol proven for a prolonged period in child language development [46, 112, 142]. M-B CDI, a standard assessment tool widely exercised in the clinical field, and its web-based version (Web-CDI) [49], employ indirect methods relying on parent reports. Recent interdisciplinary works on HCI and child language also adopt indirect methods [162, 165] as clinically favored. Overall, we designed our study with indirect assessment methods based on careful considerations of clinical advisory and methodological validity given the deployment period of 1 month.

8.3 Privacy of Home Voice Profiling

Recording and uploading real-life conversations at home would raise privacy concerns, similar to prior home-deployed systems with pervasive audio- [36, 86, 116] and video-sensing [38, 90, 92]. We expect that a careful edge-cloud separation of OSOS would mitigate much of the concerns. In our architecture (§5), most privacy-sensitive data (e.g., speech, transcription) reside in the Profiler (§5.1) and the Extractor (§5.2), which can be reasonably hosted on a home-grade device (e.g., a home server, IoT device, or even a smartphone). Recent research enabled speech-to-text recognition locally within a mobile-grade device [129, 131]. The Generator module (§5.3) would inevitably run on the cloud due to its high-end GPU requirement, but only a few target words out of the original speech survive at this stage, mitigating major privacy concerns.

8.4 Extra considerations for non-English storybooks

We employed extra engineering and human-in-the-loop steps to craft OSOS more friendly to non-English-speaking participants. It was mainly due to the GPT-4's slight underperformance in non-English languages and to make last-one-mile child-friendly touches beyond the state-of-the-art machine translation. Most human touches were put to ensure Korean-specific language styles are

---

[1] In most countries, copyright is recognized for a limited period, e.g., 70 years after the author's death in the U.S. and South Korea.

properly reflected in subtleties, e.g., many Korean words have a separate honorific form whose proper usage is clearly defined, while Korean is not as strict as English in terms of singular vs plural nouns, gender of pronouns, etc.

It is documented that English corpus dominated the trainset of GPT-3 (over 90%) [1]. Hence, we reasonably speculate that a similar language bias would exist in the GPT-4 trainset, and the generated storybooks would reflect some English-origin cultural bias. However, we did not receive feedback that children felt awkward due to apparent cultural differences, perhaps because our children are already familiar with English-origin storybooks.

Given recent breakthroughs [3, 140], rapid-developing LLMs, and diversifying the training corpus, we anticipate the extent of extra considerations may gradually shrink.

8.5 Miscellaneous Limitations and Future Work

Last one-mile automation. In designing OSOS, we intended not to fully automate the entire process for precautions in generating child-facing content and the preference for human-steerability in fine-tuning. The SLPs also favored a system reasonably interactable rather than fully autonomous. The degree of automation of a future OSOS would be an open question with the stakeholders.

Expanding the scope of language profiling. Currently, we deployed the recorders only at the participants' homes, leaving the children's other major places unprofiled, e.g., kindergarten or public spaces [83, 84, 88, 101]. It was due to the scope of consent we received. If pervasive language profiling is appreciated as a commodity educational/clinical service, the real-life coverage would grow accordingly. The SLPs called for incorporating online linguistic sources, e.g., YouTube Kids. It may require a careful license agreement review, but the technical hurdles may not be high.

Fine-grained control of the generated images. Our deployment revealed various cases calling for fine-grained control over the images, including: (1) clearly exposing the target words, (2) mitigating the Western cultural bias, and (3) fine-grained personalization such as including (excluding) the factors being positively (negatively) correlated to each child's interest. We believe that various image augmentation models such as image-to-image techniques [6, 9, 13] can steer images with precision. Moreover, a future version of OSOS may also incorporate image-to-video augmentation [11, 12, 14, 28] to further stimulate a heightened sense of interest.

Refinement of storybook interface to promote vocabulary acquisition. Literature on parent-child book reading reports explicit pointing at a picture corresponding to a target word helps the child acquire that word. On-going advances in visual-question-answering (VQA) [19] would help automate such a feature in OSOS.

9 CONCLUSION

Every child is different. So is their language environment. Different children, even if their internal language skills are equivalent, would develop nonuniform vocabulary as they are exposed to different stimuli in everyday life. Standard tools for child vocabulary assessment and intervention are inherently inflexible to individual variances. Especially these days rapidly changing and diversifying, the domain of speech-language pathology and education calls for a new complementary tool that respects one's personal differences along with their real-life context. In this paper, we presented *Open Sesame? Open Salami!* (*OSOS*), a personalized vocabulary assessment and intervention system collaboratively developed with child speech-language pathologists. Powered by generative AI and pervasive sensing, OSOS profiles a child's language environment, extracts personally tailored priority words, and generates bespoke storybooks naturally incorporating those words. Our evaluation through 4-week-long deployments to 9 families with preschool children delivered crisp findings and implications.

ACKNOWLEDGMENTS

This research was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. 2021R1A2C200386612), and by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea(NRF-2021S1A3A2A01096102). This research was also partly supported by the IITP grants by the MSIT, Korea (IITP-2020-0-01778, RS-2023-00229171).

REFERENCES

[1] 2020. GPT-3. Retrieved Dec. 6, 2023 from https://github.com/openai/gpt-3/blob/master/dataset_statistics/languages_by_word_count.csv

[2] 2021. Project Gutenberg. Retrieved Sep. 13, 2023 from https://www.gutenberg.org

[3] 2022. Machine Translation: Leaderboards. Retrieved Sep. 12, 2023 from https://paperswithcode.com/task/machine-translation

[4] 2023. CLOVA Speech Recognition (CSR). Retrieved Sep. 12, 2023 from https://www.ncloud.com/product/aiService/csr

[5] 2023. DeepL API. Retrieved Sep. 12, 2023 from https://www.deepl.com/en/docs-api

[6] 2023. Image to Image. Retrieved 2023-12-06 from https://runwayml.com/ai-magic-tools/image-to-image/

[7] 2023. KOMORAN (KOrean MORphical ANalyzer). Retrieved Sep. 12, 2023 from https://docs.komoran.kr/

[8] 2023. Library of Congress: Free to Use and Reuse: Classic Children's Books. Retrieved Sep. 13, 2023 from https://www.loc.gov/free-to-use/classic-childrens-books/

[9] 2023. Midjourney Image Prompts. Retrieved 2023-12-06 from https://docs.midjourney.com/docs/image-prompts

[10] 2023. Midjourney: Quick Start. Retrieved Sep. 12, 2023 from https://docs.midjourney.com/v1/docs

[11] 2023. Pika. Retrieved 2023-12-06 from https://pika.art/launch

[12] 2023. Runway - Advancing creativity with artificial intelligence. Retrieved 2023-12-06 from https://runwayml.com/

[13] 2023. Stability AI - Developer Platform. Retrieved 2023-12-06 from https://platform.stability.ai/docs/features/image-to-image

[14] 2023. Stable Video. Retrieved 2023-12-06 from https://stability.ai/stable-video

[15] 2023. The World's Largest Profanity List. Retrieved Sep. 12, 2023 from https://www.surgehq.ai/datasets/profanity-dataset

[16] Arwa I Alhussain and Aqil M Azmi. 2021. Automatic story generation: a survey of approaches. *ACM Computing Surveys (CSUR)* 54, 5 (2021), 1–38.

[17] Nina J Anderson, Susan A Graham, Heather Prime, Jennifer M Jenkins, and Sheri Madigan. 2021. Linking quality and quantity of parental linguistic input to child language skills: A meta-analysis. *Child Development* 92, 2 (2021), 484–501.

[18] Jeremy M Anglin, George A Miller, and Pamela C Wakefield. 1993. Vocabulary development: A morphological analysis. *Monographs of the society for research in child development* (1993), i–186.

[19] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*. 2425–2433.

[20] Alan Baddeley. 2003. Working memory and language: An overview. *Journal of communication disorders* 36, 3 (2003), 189–208.

[21] David Barker. 2007. A personalized approach to analyzing 'cost'and 'benefit'in vocabulary selection. *System* 35, 4 (2007), 523–533.

[22] Elizabeth Bates and Brian MacWhinney. 1987. Competition, variation, and language learning. (1987).

[23] Elizabeth Bates and Brian MacWhinney. 2014. Competition, variation, and language learning. In *Mechanisms of language acquisition*. Psychology Press, 157–193.

[24] Elizabeth Bates, Virginia Marchman, Donna Thal, Larry Fenson, Philip Dale, J Steven Reznick, Judy Reilly, and Jeff Hartung. 1994. Developmental and stylistic variation in the composition of early vocabulary. *Journal of child language* 21, 1 (1994), 85–123.

[25] Isabel L Beck, Margaret G McKeown, and Linda Kucan. 2013. *Bringing words to life: Robust vocabulary instruction.* Guilford Press.

[26] Joseph H Beitchman, Beth Wilson, Carla J Johnson, Leslie Atkinson, Arlene Young, Edward Adlaf, Michael Escobar, and Lori Douglas. 2001. Fourteen-year follow-up of speech/language-impaired and control children: Psychiatric outcome. *Journal of the American Academy of Child & Adolescent Psychiatry* 40, 1 (2001), 75–82.

[27] John B Black and Robert Wilensky. 1979. An evaluation of story grammars. *Cognitive science* 3, 3 (1979), 213–229.

[28] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. 2023. Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets. *arXiv preprint arXiv:2311.15127* (2023).

[29] Nikoletta Bogka, Jackie Masterson, Judit Druks, Meltini Fragkioudaki, Eftychia-Sapfo Chatziprokopiou, and Konstantinos Economou. 2003. Object and action picture naming in English and Greek. *European Journal of Cognitive Psychology* 15, 3 (2003), 371–403.

[30] Amy Burger and Ivy Chong. 2011. *Receptive Vocabulary.* Springer US, Boston, MA, 1231–1231. https://doi.org/10.1007/978-0-387-79061-9_2359

[31] Xiang Cao, Siân E Lindley, John Helmes, and Abigail Sellen. 2010. Telling the whole story: anticipation, inspiration and reputation in a field deployment of TellTable. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work.* 251–260.

[32] Susan Carey and Elsa Bartlett. 1978. Acquiring a single new word. (1978).

[33] Hugh W Catts and Alan G Kamhi. 2005. *The connections between language and reading disabilities.* Psychology Press.

[34] F Cecchi, SM Serio, M Del Maestro, C Laschi, and P Dario. 2010. Design and development of sensorized toys for monitoring infants' grasping actions. In *2010 3rd IEEE RAS & EMBS International Conference on Biomedical Robotics and Biomechatronics.* IEEE, 247–252.

[35] Jae-Eun Cha, Jung-Mee Kim, Soo-Jin Kim, Mi-Sun Yoon, and Moon-Soo Chang. 2014. Substantives in the Vocabulary of Typically Developing Young Children. *Communication Sciences & Disorders* 19, 4 (2014), 430–446.

[36] Narae Cha, Auk Kim, Cheul Young Park, Soowon Kang, Mingyu Park, Jae-Gil Lee, Sangsu Lee, and Uichin Lee. 2020. Hello there! is now a good time to talk? Opportune moments for proactive interactions with smart speakers. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 3 (2020), 1–28.

[37] Tiffany Chen, Cassandra Lee, Jessica R Mindel, Neska Elhaouij, and Rosalind Picard. 2023. Closer Worlds: Using Generative AI to Facilitate Intimate Conversations. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems.* 1–15.

[38] Sungjae Cho, Yoonsu Kim, Jaewoong Jang, and Inseok Hwang. 2023. AI-to-Human Actuation: Boosting Unmodified AI's Robustness by Proactively Inducing Favorable Human Sensing Conditions. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 1 (2023), 1–32.

[39] B Choe, I-h Lee, and S-g Lee. 2020. Korean morphological analyzer for neologism and spacing error based on sequence-to-sequence. *J KIISE* 47 (2020), 70–77.

[40] YunSeok Choi, SuAh Kim, and Jee-Hyong Lee. 2016. Recurrent neural network for storytelling. In *2016 Joint 8th international conference on soft computing and intelligent systems (SCIS) and 17th international symposium on advanced intelligent systems (ISIS).* IEEE, 841–845.

[41] 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 *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems.* 1–19.

[42] Kenneth Ward Church. 2017. Word2Vec. *Natural Language Engineering* 23, 1 (2017), 155–162.

[43] Judy Clegg, Chris Hollis, Lynn Mawhood, and Michael Rutter. 2005. Developmental language disorders–a follow-up in later adult life. Cognitive, language and psychosocial outcomes. *Journal of child psychology and psychiatry* 46, 2 (2005), 128–149.

[44] Cynthia G Clopper. 2021. Perception of dialect variation. *The handbook of speech perception* (2021), 333–364.

[45] Bernard Comrie. 1988. Linguistic typology. *Annual Review of Anthropology* 17, 1 (1988), 145–159.

[46] Philip S Dale. 1991. The validity of a parent report measure of vocabulary and syntax at 24 months. *Journal of Speech, Language, and Hearing Research* 34, 3 (1991), 565–571.

[47] Lilian De Greef, Mayank Goel, Min Joon Seo, Eric C Larson, James W Stout, James A Taylor, and Shwetak N Patel. 2014. Bilicam: using mobile phones to monitor newborn jaundice. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing.* 331–342.

[48] Annick De Houwer, Marc H Bornstein, and Diane L Putnick. 2014. A bilingual–monolingual comparison of young children's vocabulary size: Evidence from comprehension and production. *Applied psycholinguistics* 35, 6 (2014), 1189–1211.

[49] Benjamin DeMayo, Danielle Kellier, Mika Braginsky, Christina Bergmann, Cielke Hendriks, Caroline F Rowland, Michael Frank, and Virginia Marchman. 2021. Web-CDI: A system for online administration of the MacArthur-Bates Communicative Development Inventories. *Language Development Research* (2021).

[50] Nicoletta Di Blas and Bianca Boretti. 2009. Interactive storytelling in pre-school: a case-study. In *Proceedings of the 8th International conference on interaction design and children.* 44–51.

[51] DM Dunn. 2018. Peabody picture vocabulary test fifth edition (PPVT-5). *Minneapolis, MN* (2018).

[52] Elisabeth Duursma, Marilyn Augustyn, and Barry Zuckerman. 2008. Reading aloud to children: the evidence. *Archives of disease in childhood* 93, 7 (2008), 554–557.

[53] Susanne Ebert, Kathrin Lockl, Sabine Weinert, Yvonne Anders, Katharina Kluczniok, and Hans-Günther Rossbach. 2013. Internal and external influences on vocabulary development in preschool children. *School Effectiveness and School Improvement* 24, 2 (2013), 138–154.

[54] Franziska Egert, Anne-Kristin Cordes, and Fabienne Hartig. 2022. Can e-books foster child language? Meta-analysis on the effectiveness of e-book interventions in early childhood education and care. *Educational Research Review* 37 (2022), 100472.

[55] Amy M Elleman, Endia J Lindo, Paul Morphy, and Donald L Compton. 2009. The impact of vocabulary instruction on passage-level comprehension of school-age children: A meta-analysis. *Journal of Research on Educational Effectiveness* 2, 1 (2009), 1–44.

[56] Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833* (2018).

[57] Larry Fenson et al. 2007. MacArthur-Bates communicative development inventories. (2007).

[58] Larry Fenson, Philip S Dale, J Steven Reznick, Elizabeth Bates, Donna J Thal, Stephen J Pethick, Michael Tomasello, Carolyn B Mervis, and Joan Stiles. 1994. Variability in early communicative development. *Monographs of the society for research in child development* (1994), i–185.

[59] Larry Fenson, Steve Pethick, Connie Renda, Jeffrey L Cox, Philip S Dale, and J Steven Reznick. 2000. Short-form versions of the MacArthur communicative development inventories. *Applied psycholinguistics* 21, 1 (2000), 95–116.

[60] Anne Fernald, Virginia A Marchman, and Adriana Weisleder. 2013. SES differences in language processing skill and vocabulary are evident at 18 months. *Developmental science* 16, 2 (2013), 234–248.

[61] Sean Follmer, Rafael Ballagas, Hayes Raffle, Mirjana Spasojevic, and Hiroshi Ishii. 2012. People in books: using a FlashCam to become part of an interactive book for connected reading. In *Proceedings of the ACM 2012 conference on Computer supported cooperative work.* 685–694.

[62] William K Frankenburg, Josiah Dodds, Philip Archer, Howard Shapiro, and Beverly Bresnick. 1992. The Denver II: a major revision and restandardization of the Denver Developmental Screening Test. *Pediatrics* 89, 1 (1992), 91–97.

[63] Susan E Gathercole, Catherine S Willis, Hazel Emslie, and Alan D Baddeley. 1992. Phonological memory and vocabulary development during the early school years: A longitudinal study. *Developmental psychology* 28, 5 (1992), 887.

[64] Dedre Gentner. 1982. Why nouns are learned before verbs: Linguistic relativity versus natural partitioning. *BBN report; no. 4854* (1982).

[65] Jill Gilkerson, Jeffrey A Richards, Steven F Warren, Judith K Montgomery, Charles R Greenwood, D Kimbrough Oller, John HL Hansen, and Terrance D Paul. 2017. Mapping the early language environment using all-day recordings and automated analysis. *American journal of speech-language pathology* 26, 2 (2017), 248–265.

[66] Jennifer M Goldschneider and Robert M DeKeyser. 2001. Explaining the "natural order of L2 morpheme acquisition" in English: A meta-analysis of multiple determinants. *Language learning* 51, 1 (2001), 1–50.

[67] Yuan Gong, Youxin Pang, Xiaodong Cun, Menghan Xia, Haoxin Chen, Longyue Wang, Yong Zhang, Xintao Wang, Ying Shan, and Yujiu Yang. 2023. TaleCrafter: Interactive Story Visualization with Multiple Characters. *arXiv preprint arXiv:2305.18247* (2023).

[68] Judith C Goodman, Philip S Dale, and Ping Li. 2008. Does frequency count? Parental input and the acquisition of vocabulary. *Journal of child language* 35, 3 (2008), 515–531.

[69] Joseph H Greenberg. 1960. A quantitative approach to the morphological typology of language. *International journal of American linguistics* 26, 3 (1960), 178–194.

[70] Elizabeth B Hadley, David K Dickinson, Kathy Hirsh-Pasek, and Roberta Michnick Golinkoff. 2019. Building semantic networks: The impact of a vocabulary intervention on preschoolers' depth of word knowledge. *Reading Research Quarterly* 54, 1 (2019), 41–61.

[71] Ariel Han and Zhenyao Cai. 2023. Design implications of generative AI systems for visual storytelling for young learners. In *Proceedings of the 22nd Annual*

ACM Interaction Design and Children Conference. 470–474.

[72] Brent Harrison, Christopher Purdy, and Mark Riedl. 2017. Toward automated story generation with markov chain monte carlo methods and deep neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, Vol. 13. 191–197.

[73] Betty Hart and Todd R Risley. 2003. The early catastrophe: The 30 million word gap by age 3. *American educator* 27, 1 (2003), 4–9.

[74] Betty Hart, Todd R Risley, and John R Kirby. 1997. Meaningful differences in the everyday experience of young American children. *Canadian Journal of Education* 22, 3 (1997), 323.

[75] Ian Hay, Gordon Elias, Ruth Fielding-Barnsley, Ross Homel, and Kate Freiberg. 2007. Language delays, reading delays, and learning difficulties: Interactive elements requiring multidimensional programming. *Journal of Learning Disabilities* 40, 5 (2007), 400–409.

[76] Elfrieda H Hiebert, Judith A Scott, Ruben Castaneda, and Alexandra Spichtig. 2019. An analysis of the features of words that influence vocabulary difficulty. *Education Sciences* 9, 1 (2019), 8.

[77] Erika Hoff. 2003. The specificity of environmental influence: Socioeconomic status affects early vocabulary development via maternal speech. *Child development* 74, 5 (2003), 1368–1378.

[78] Erika Hoff. 2013. *Language development*. Cengage Learning.

[79] Erika Hoff and Letitia Naigles. 2002. How children use input to acquire a lexicon. *Child development* 73, 2 (2002), 418–433.

[80] Erika Hoff, Rosario Rumiche, Andrea Burridge, Krystal M Ribot, and Stephanie N Welsh. 2014. Expressive vocabulary development in children from bilingual and monolingual homes: A longitudinal study from two to four years. *Early childhood research quarterly* 29, 4 (2014), 433–444.

[81] Sarah McCue Horwitz, Julia R Irwin, Margaret J Briggs-Gowan, Joan M Bosson Heenan, Jennifer Mendoza, and Alice S Carter. 2003. Language delay in a community cohort of young children. *Journal of the American Academy of Child & Adolescent Psychiatry* 42, 8 (2003), 932–940.

[82] Chieh-Yang Huang, Shih-Hong Huang, and Ting-Hao Kenneth Huang. 2020. Heteroglossia: In-situ story ideation with the crowd. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[83] Inseok Hwang, Hyukjae Jang, Lama Nachman, and Junehwa Song. 2010. Exploring inter-child behavioral relativity in a shared social environment: a field study in a kindergarten. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*. 271–280.

[84] Inseok Hwang, Hyukjae Jang, Taiwoo Park, Aram Choi, Youngki Lee, Chanyou Hwang, Yanggui Choi, Lama Nachman, and Junehwa Song. 2012. Leveraging children's behavioral distribution and singularities in new interactive environments: Study in kindergarten field trips. In *Pervasive Computing: 10th International Conference, Pervasive 2012, Newcastle, UK, June 18-22, 2012. Proceedings 10*. Springer, 39–56.

[85] Inseok Hwang, Youngki Lee, Chungkuk Yoo, Chulhong Min, Dongsun Yim, and John Kim. 2019. Towards interpersonal assistants: next-generation conversational agents. *IEEE Pervasive Computing* 18, 2 (2019), 21–31.

[86] Inseok Hwang, Chungkuk Yoo, Chanyou Hwang, Dongsun Yim, Youngki Lee, Chulhong Min, John Kim, and Junehwa Song. 2014. TalkBetter: family-driven mobile intervention care for children with language delay. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. 1283–1296.

[87] Erin M Ingvalson, Lynn K Perry, Mark VanDam, and Tina M Grieco-Calub. 2023. Comparing Scores on the Peabody Picture Vocabulary Test and Receptive One-Word Picture Vocabulary Test in Preschoolers With and Without Hearing Loss. *American journal of speech-language pathology* (2023), 1–10.

[88] Hyukjae Jang, Sungwon Peter Choe, Inseok Hwang, Chanyou Hwang, Lama Nachman, and Junehwa Song. 2012. RubberBand: augmenting teacher's awareness of spatially isolated children on kindergarten field trips. In *Proceedings of the 2012 ACM conference on ubiquitous computing*. 236–239.

[89] Jasmine Jones, David Merritt, and Mark S Ackerman. 2017. KidKeeper: Design for capturing audio mementos of everyday life for parents of young children. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 1864–1875.

[90] Bumsoo Kang, Inseok Hwang, Jinho Lee, Seungchul Lee, Taegyeong Lee, Youngjae Chang, and Min Kyung Lee. 2018. My being to your place, your being to my place: Co-present robotic avatars create illusion of living together. In *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services*. 54–67.

[91] Bumsoo Kang, Seungwoo Kang, and Inseok Hwang. 2021. Momentmeld: Ai-augmented mobile photographic memento towards mutually stimulatory intergenerational interaction. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[92] Bumsoo Kang, Seungwoo Kang, and Inseok Hwang. 2023. AI-driven Family Interaction Over Melded Space and Time. *IEEE Pervasive Computing* 22, 1 (2023), 85–94.

[93] Bumsoo Kang, Chulhong Min, Wonjung Kim, Inseok Hwang, Chunjong Park, Seungchul Lee, Sung-Ju Lee, and Junehwa Song. 2017. Zaturi: We put together the 25th hour for you. create a book for your baby. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 1850–1863.

[94] Seokbin Kang, Youngwoon Lee, and Suwoong Lee. 2015. Kids in Fairytales: Experiential and Interactive Storytelling in Children's Libraries. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*. 1007–1012.

[95] Elizabeth Kelley, Garland Jones, and Deborah Fein. 2003. Language assessment in children. *Comprehensive handbook of psychological assessment: Intellectual and neuropsychological assessment* 1 (2003), 191–215.

[96] Julie A Kientz, Rosa I Arriaga, and Gregory D Abowd. 2009. Baby steps: evaluation of a system to support record-keeping for parents of young children. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1713–1722.

[97] Ho-hyun Kil. 2018. The study of Korean stopwords list for text mining. *URIMALGEUL: The Korean Language and Literature* 78 (2018), 1–25.

[98] Eunji Kim and Gweeyeon Jeon. 2020. Effect of the use of smart media on the cognitive and language development of the preschooler: The mediating effect of smart media addiction tendency and the moderated mediating effects of maternal guidance on smart media usage. *Family and Environment Research* 58, 1 (2020), 13–29.

[99] Joy Kim, Justin Cheng, and Michael S Bernstein. 2014. Ensemble: exploring complementary strengths of leaders and crowds in creative collaboration. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. 745–755.

[100] Joy Kim, Sarah Sterman, Allegra Argent Beal Cohen, and Michael S Bernstein. 2017. Mechanical novel: Crowdsourcing complex work through reflection and revision. In *Proceedings of the 2017 acm conference on computer supported cooperative work and social computing*. 233–245.

[101] Wonjung Kim, Seungchul Lee, Youngjae Chang, Taegyeong Lee, Inseok Hwang, and Junehwa Song. 2021. Hivemind: social control-and-use of IoT towards democratization of public spaces. In *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services*. 467–482.

[102] Wonjung Kim, Seungchul Lee, Seonghoon Kim, Sungbin Jo, Chungkuk Yoo, Inseok Hwang, Seungwoo Kang, and Junehwa Song. 2020. Dyadic Mirror: Everyday Second-person Live-view for Empathetic Reflection upon Parent-child Interaction. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 3 (2020), 1–29.

[103] Young Tae Kim, Gyung Hun Hong, and Kyung Hee Kim. 2009. Content and reliability analyses of the receptive and expressive vocabulary test (REVT). *Communication Sciences & Disorders* 14, 1 (2009), 34–45.

[104] Natalia Kucirkova. 2019. Children's agency by design: Design parameters for personalization in story-making apps. *International Journal of Child-Computer Interaction* 21 (2019), 112–120.

[105] Natalia Kucirkova, Merideth Gattis, Thomas P Spargo, Beatriz Seisdedos de Vega, and Rosie Flewitt. 2021. An empirical investigation of parent-child shared reading of digital personalized books. *International Journal of Educational Research* 105 (2021), 101710.

[106] Natalia Kucirkova, David Messer, and Kieron Sheehy. 2014. Reading personalized books with preschool children enhances their word acquisition. *First Language* 34, 3 (2014), 227–243.

[107] Natalia Kucirkova, David Messer, and Denise Whitelock. 2013. Parents reading with their toddlers: The role of personalization in book engagement. *Journal of Early Childhood Literacy* 13, 4 (2013), 445–470.

[108] Taeahn Kwon, Minkyung Jeong, Eon-Suk Ko, and Youngki Lee. 2022. Captivate! contextual language guidance for parent–child interaction. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–17.

[109] George Lakoff. 1972. Structural complexity in fairy tales. (1972).

[110] Eric C Larson, Mayank Goel, Gaetano Boriello, Sonya Heltshe, Margaret Rosenfeld, and Shwetak N Patel. 2012. SpiroSmart: using a microphone to measure lung function on a mobile phone. In *Proceedings of the 2012 ACM Conference on ubiquitous computing*. 280–289.

[111] Emily Laubscher and Janice Light. 2020. Core vocabulary lists for young children and considerations for early language development: A narrative review. *Augmentative and Alternative Communication* 36, 1 (2020), 43–53.

[112] James Law and Penny Roy. 2008. Parental report of infant language skills: A review of the development and application of the Communicative Development Inventories. *Child and adolescent mental health* 13, 4 (2008), 198–206.

[113] Jungeun Lee and Inseok Hwang. 2023. Hyper-personalizing Common Norms through Principled Bespoke Generation. In *Adjunct Proceedings of the 2023 ACM International Joint Conference on Pervasive and Ubiquitous Computing & the 2023 ACM International Symposium on Wearable Computing*. 215–219.

[114] Jungeun Lee, Sungnam Kim, Minki Cheon, Hyojin Ju, JaeEun Lee, and Inseok Hwang. 2022. SleepGuru: Personalized Sleep Planning System for Real-life Actionability and Negotiability. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–16.

[115] Yoonjoo Lee, Tae Soo Kim, Minsuk Chang, and Juho Kim. 2022. Interactive Children's Story Rewriting Through Parent-Children Interaction. In *Proceedings*

of the First Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2022). 62–71.

[116] Youngki Lee, Chulhong Min, Chanyou Hwang, Jaeung Lee, Inseok Hwang, Younghyun Ju, Chungkuk Yoo, Miri Moon, Uichin Lee, and Junehwa Song. 2013. Sociophone: Everyday face-to-face interaction monitoring platform using multi-phone sensor fusion. In *Proceeding of the 11th annual international conference on Mobile systems, applications, and services*. 375–388.

[117] Geoffrey Leech, Paul Rayson, et al. 2014. *Word frequencies in written and spoken English: Based on the British National Corpus*. Routledge.

[118] Boyang Li, Stephen Lee-Urban, George Johnston, and Mark Riedl. 2013. Story generation with crowdsourced plot graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 27. 598–604.

[119] Vivian Liu and Lydia B Chilton. 2022. Design guidelines for prompt engineering text-to-image generative models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–23.

[120] Michelle Macroy-Higgins and Elizabeth A Montemarano. 2016. Attention and word learning in toddlers who are late talkers. *Journal of child language* 43, 5 (2016), 1020–1037.

[121] Sheri Madigan, Heather Prime, Susan A Graham, Michelle Rodrigues, Nina Anderson, Jennifer Khoury, and Jennifer M Jenkins. 2019. Parenting behavior and child language: A meta-analysis. *Pediatrics* 144, 4 (2019).

[122] Adyasha Maharana, Darryl Hannan, and Mohit Bansal. 2022. Storydall-e: Adapting pretrained text-to-image transformers for story continuation. In *European Conference on Computer Vision*. Springer, 70–87.

[123] Christopher D Manning. 2011. Part-of-speech tagging from 97% to 100%: is it time for some linguistics?. In *International conference on intelligent text processing and computational linguistics*. Springer, 171–189.

[124] James Manyika. 2023. *An overview of Bard: an early experiment with generative AI*. Technical Report. Technical report, Google AI.

[125] Virginia A Marchman and Philip S Dale. 2023. The MacArthur-Bates Communicative Development Inventories: updates from the CDI Advisory Board. *Frontiers in Psychology* 14 (2023), 1170303.

[126] Clara D Martin, Xavier Garcia, Douglas Potter, Alissa Melinger, and Albert Costa. 2016. Holiday or vacation? The processing of variation in vocabulary across dialects. *Language, Cognition and Neuroscience* 31, 3 (2016), 375–390.

[127] April S Masarik and Rand D Conger. 2017. Stress and child development: A review of the Family Stress Model. *Current opinion in psychology* 13 (2017), 85–90.

[128] Sandra McCormick. 1977. Should you read aloud to your children? *Language Arts* 54, 2 (1977), 139–163.

[129] Ian McGraw, Rohit Prabhavalkar, Raziel Alvarez, Montse Gonzalez Arenas, Kanishka Rao, David Rybach, Ouais Alsharif, Haşim Sak, Alexander Gruenstein, Françoise Beaufays, et al. 2016. Personalized speech recognition on mobile devices. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5955–5959.

[130] Suzanne E Mol, Adriana G Bus, and Maria T De Jong. 2009. Interactive book reading in early education: A tool to stimulate print knowledge as well as oral language. *Review of Educational Research* 79, 2 (2009), 979–1007.

[131] Mohammed Kyari Mustafa, Tony Allen, and Kofi Appiah. 2019. A comparative review of dynamic neural networks and hidden Markov model methods for mobile on-device speech recognition. *Neural Computing and Applications* 31 (2019), 891–899.

[132] Susan B Neuman, Ellen H Newman, and Julie Dwyer. 2011. Educational effects of a vocabulary intervention on preschoolers' word knowledge and conceptual development: A cluster-randomized trial. *Reading Research Quarterly* 46, 3 (2011), 249–272.

[133] Maria Nikolajeva. 2003. Verbal and visual literacy: the role of picturebooks in the reading experience of young children. *Handbook of early childhood literacy* (2003), 235–248.

[134] Claire Noble, Giovanni Sala, Michelle Peter, Jamie Lingwood, Caroline Rowland, Fernand Gobet, and Julian Pine. 2019. The impact of shared book reading on children's language skills: A meta-analysis. *Educational Research Review* 28 (2019), 100290.

[135] Jea-Hyuk Oh, Jae-Eun Cha, Mi-Sun Yoon, Jung-Mee Kim, and Moon-Soo Chang. 2014. Predicates in the Vocabulary of Typically Developing Preschoolers. *Communication Sciences & Disorders* 19, 3 (2014), 274–284.

[136] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]

[137] Hiroyuki Osone, Jun-Li Lu, and Yoichi Ochiai. 2021. BunCho: ai supported story co-creation via unsupervised multitask learning to increase writers' creativity in japanese. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–10.

[138] Amy Pace, Rufan Luo, Kathy Hirsh-Pasek, and Roberta Michnick Golinkoff. 2017. Identifying pathways between socioeconomic status and language development. *Annual Review of Linguistics* 3 (2017), 285–308.

[139] S Pae and K Kwak. 2011. Korean MacArthur-Bates Communicative Development Inventories User's Guide and Technical Manual. *Seoul: Mindpress Publishing Co* (2011).

[140] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 311–318.

[141] Johanne Paradis. 2011. Individual differences in child English second language acquisition: Comparing child-internal and child-external factors. *Linguistic approaches to bilingualism* 1, 3 (2011), 213–237.

[142] Johanne Paradis, Kristyn Emmerzael, and Tamara Sorenson Duncan. 2010. Assessment of English language learners: Using parent report on first language development. *Journal of communication disorders* 43, 6 (2010), 474–497.

[143] Rhea Paul. 2007. *Language disorders from infancy through adolescence: Assessment & intervention*. Vol. 324. Elsevier Health Sciences.

[144] Kim Plunkett. 1993. Lexical segmentation and vocabulary growth in early language acquisition. *Journal of Child Language* 20, 1 (1993), 43–60.

[145] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. arXiv:2307.01952 [cs.CV]

[146] Hayes Raffle, Rafael Ballagas, Glenda Revelle, Hiroshi Horii, Sean Follmer, Janet Go, Emily Reardon, Koichi Mori, Joseph Kaye, and Mirjana Spasojevic. 2010. Family story play: reading with young children (and elmo) over a distance. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1583–1592.

[147] Hayes Raffle, Glenda Revelle, Koichi Mori, Rafael Ballagas, Kyle Buza, Hiroshi Horii, Joseph Kaye, Kristin Cook, Natalie Freed, Janet Go, et al. 2011. Hello, is grandma there? let's read! StoryVisit: family video chat and connected e-books. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1195–1204.

[148] Tanzila Rahman, Hsin-Ying Lee, Jian Ren, Sergey Tulyakov, Shweta Mahajan, and Leonid Sigal. 2023. Make-a-story: Visual memory conditioned consistent story generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2493–2502.

[149] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*. PMLR, 8821–8831.

[150] Hannah Rashkin, Asli Celikyilmaz, Yejin Choi, and Jianfeng Gao. 2020. Plotmachines: Outline-conditioned generation with dynamic plot state tracking. *arXiv preprint arXiv:2004.14967* (2020).

[151] Alvin E Roth and Elliott Peranson. 1999. The redesign of the matching market for American physicians: Some engineering aspects of economic design. *American economic review* 89, 4 (1999), 748–780.

[152] Brandon Cain Roy and Deb K Roy. 2009. Fast transcription of unstructured audio recordings. (2009).

[153] Elisa Rubegni and Monica Landoni. 2014. Fiabot! Design and evaluation of a mobile storytelling application for schools. In *Proceedings of the 2014 conference on Interaction design and children*. 165–174.

[154] Elisa Rubegni and Monica Landoni. 2015. Supporting creativity in designing story authoring tools. In *Proceedings of the 14th International Conference on Interaction Design and Children*. 287–290.

[155] Elisa Rubegni and Monica Landoni. 2018. How to design a digital storytelling authoring tool for developing pre-reading and pre-writing skills. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–10.

[156] James Ryan. 2017. Grimes' fairy tales: a 1960s story generator. In *Interactive Storytelling: 10th International Conference on Interactive Digital Storytelling, ICIDS 2017 Funchal, Madeira, Portugal, November 14–17, 2017, Proceedings 10*. Springer, 89–103.

[157] Steffi Sachse and Waldemar Von Suchodoletz. 2008. Early identification of language delay by direct language assessment or parent report? *Journal of Developmental & Behavioral Pediatrics* 29, 1 (2008), 34–41.

[158] Paula J Schwanenflugel. 2013. Why are abstract concepts hard to understand? In *The psychology of word meanings*. Psychology Press, 235–262.

[159] Monique Sénéchal. 1997. The differential effect of storybook reading on preschoolers' acquisition of expressive and receptive vocabulary. *Journal of child language* 24, 1 (1997), 123–138.

[160] Hanieh Shakeri, Carman Neustaedter, and Steve DiPaola. 2021. Saga: Collaborative storytelling with gpt-3. In *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing*. 163–166.

[161] MAQA Shareah, BAD Mudhsh, and Ayman Hamid AL-Takhayinh. 2015. An overview on dialectal variation. *International Journal of Scientific and Research Publications* 5, 6 (2015), 1–5.

[162] Donghoon Shin, Jaeyoon Song, Seokwoo Song, Jisoo Park, Joonhwan Lee, and Soojin Jun. 2020. TalkingBoogie: collaborative mobile AAC system for non-verbal children with developmental disabilities and their caregivers. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[163] Lei Shu, Liangchen Luo, Jayakumar Hoskere, Yun Zhu, Canoee Liu, Simon Tong, Jindong Chen, and Lei Meng. 2023. RewriteLM: An Instruction-Tuned Large Language Model for Text Rewriting. *arXiv preprint arXiv:2305.15685* (2023).

[164] Ho-Min Sohn. 2001. *The korean language*. Cambridge University Press.

[165] Seokwoo Song, Juho Kim, Bumsoo Kang, Wonjeong Park, and John Kim. 2018. BebeCODE: Collaborative Child Development Tracking System. In *Proceedings of the 2018 chi conference on human factors in computing systems*. 1–12.

[166] Seokwoo Song, Seungho Kim, John Kim, Wonjeong Park, and Dongsun Yim. 2016. TalkLIME: mobile system intervention to improve parent-child interaction for children with language delay. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 304–315.

[167] Laura M Steacy, Douglas Fuchs, Jennifer K Gilbert, Devin M Kearns, Amy M Elleman, and Ashley A Edwards. 2020. Sight word acquisition in first grade students at risk for reading disabilities: An item-level exploration of the number of exposures required for mastery. *Annals of dyslexia* 70 (2020), 259–274.

[168] Laura M Steacy, Devin M Kearns, Jennifer K Gilbert, Donald L Compton, Eunsoo Cho, Esther R Lindstrom, and Alyson A Collins. 2017. Exploring individual differences in irregular word recognition among children with early-emerging and late-emerging word reading difficulty. *Journal of Educational Psychology* 109, 1 (2017), 51.

[169] Holly L Storkel, Rouzana Komesidou, Kandace K Fleming, and Rebecca Swinburne Romine. 2017. Interactive book reading to accelerate word learning by kindergarten children with specific language impairment: Identifying adequate progress and successful learning patterns. *Language, speech, and hearing services in schools* 48, 2 (2017), 108–124.

[170] Anselm Strauss and Juliet Corbin. 1998. Basics of qualitative research techniques. (1998).

[171] Hyewon Suh, John R Porter, Alexis Hiniker, and Julie A Kientz. 2014. @ BabySteps: design and evaluation of a system for using twitter for tracking children's developmental milestones. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2279–2288.

[172] Hyewon Suh, John R Porter, Robert Racadio, Yi-Chen Sung, and Julie A Kientz. 2016. Baby steps text: Feasibility study of an SMS-based tool for tracking children's developmental progress. In *AMIA Annual Symposium Proceedings*, Vol. 2016. American Medical Informatics Association, 1997.

[173] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

[174] Jacqueline A Towson, Yusuf Akemoglu, Laci Watkins, and Songtian Zeng. 2021. Shared interactive book reading interventions for young children with disabilities: A systematic review. *American Journal of Speech-Language Pathology* 30, 6 (2021), 2700–2715.

[175] Ann A Tyler and Leslie C Tolbert. 2002. Speech-language assessment in the clinical setting. (2002).

[176] Sveva Valguarnera and Monica Landoni. 2022. ROBIN-Designing a ROBot for Interactive Narratives to engage preschool children. In *6th FabLearn Europe/MakeEd Conference 2022*. 1–5.

[177] Sveva Valguarnera and Monica Landoni. 2022. " This book is magical!": exploring emergent readers' preferences and wishes for storytelling tools. In *Nordic Human-Computer Interaction Conference*. 1–9.

[178] Ellen Vandewalle, Bart Boets, Pol Ghesquière, and Inge Zink. 2012. Development of phonological processing skills in children with specific language impairment with and without literacy delay: A 3-year longitudinal study. (2012).

[179] Louiza Voniati, Spyros Armostis, and Dionysios Tafiadis. 2021. Language sampling practices: A review for clinicians. *Evidence-Based Communication Assessment and Intervention* 15, 1 (2021), 24–45.

[180] Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T Campbell. 2014. StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing*. 3–14.

[181] Barbara A Wasik, Annemarie H Hindman, and Emily K Snell. 2016. Book reading and vocabulary development: A systematic review. *Early Childhood Research Quarterly* 37 (2016), 39–57.

[182] Tracy L Westeyn, Gregory D Abowd, Thad E Starner, Jeremy M Johnson, Peter W Presti, and Kimberly A Weaver. 2012. Monitoring children's developmental progress using augmented toys and activity recognition. *Personal and Ubiquitous Computing* 16 (2012), 169–191.

[183] Grover J Whitehurst, Francine L Falco, Christopher J Lonigan, Janet E Fischel, Barbara D DeBaryshe, Marta C Valdez-Menchaca, and Marie Caulfield. 1988. Accelerating language development through picture book reading. *Developmental psychology* 24, 4 (1988), 552.

[184] Kathleen T Williams. 1997. Expressive vocabulary test second edition (EVT™ 2). *J. Am. Acad. Child Adolesc. Psychiatry* 42 (1997), 864–872.

[185] Stefanie Wulff and Nick C Ellis. 2018. Usage-based approaches to second language acquisition. I: David T. Miller, Faith Bayram, Jason Rothman & Ludovica Serratrice (red.): Bilingual cognition and language. The state of the science across its subfields.

[186] 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.

[187] Shunyu Yao, Howard Chen, Austin W Hanjie, Runzhe Yang, and Karthik Narasimhan. 2023. Collie: Systematic construction of constrained text generation tasks. *arXiv preprint arXiv:2307.08689* (2023).

[188] Dongsun Yim, Wonjeong Park, Seonghye Cheon, Yeo-Jin Lee, and Jiyeon Lee. 2015. Interaction skills via spontaneous speech sample analysis: 2-to 5-year-old children with and without language impairment. *Communication Sciences & Disorders* 20, 3 (2015), 364–373.

[189] Chungkuk Yoo, Seungwoo Kang, Inseok Hwang, Chulhong Min, Seonghoon Kim, Wonjung Kim, and Junehwa Song. 2019. Mom, I see You Angry at Me! Designing a Mobile Service for Parent-child Conflicts by In-situ Emotional Empathy. In *Proceedings of the 5th ACM Workshop on Mobile Systems for Computational Social Science*. 21–26.

[190] Christine Yoshinaga-Itano. 2006. Early identification, communication modality, and the development of speech and spoken language skills: Patterns and considerations. *Advances in the spoken language development of deaf and hard-of-hearing children* (2006), 298–327.

[191] Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: story writing with large language models. In *27th International Conference on Intelligent User Interfaces*. 841–852.

[192] Niloofar Zarei, Sharon Lynn Chu, Francis Quek, Nanjie'Jimmy' Rao, and Sarah Anne Brown. 2020. Investigating the Effects of Self-Avatars and Story-Relevant Avatars on Children's Creative Storytelling. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–11.

[193] Andrea A Zevenbergen and Grover J Whitehurst. 2003. Dialogic reading: A shared picture book reading intervention for preschoolers. *On reading books to children: Parents and teachers* 177 (2003), 200.

[194] 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.

[195] 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.

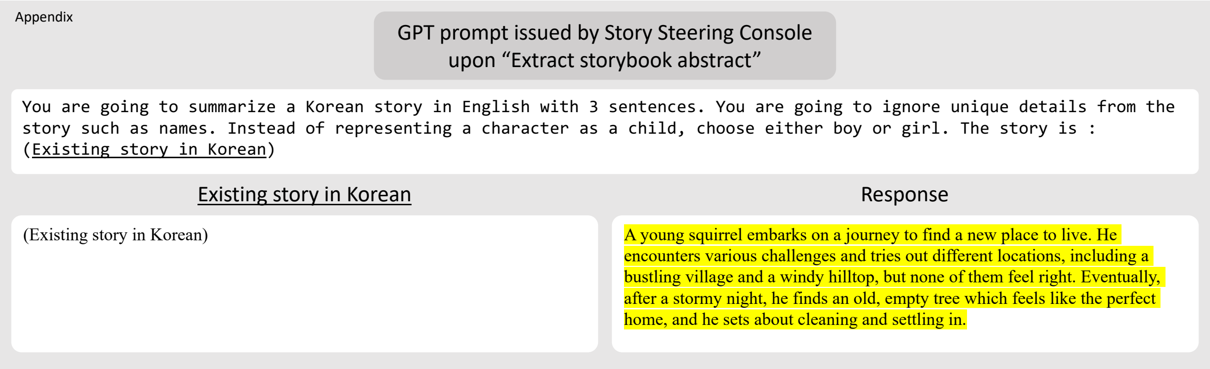
[196] Wangchunshu Zhou, Yuchen Eleanor Jiang, Ethan Wilcox, Ryan Cotterell, and Mrinmaya Sachan. 2023. Controlled text generation with natural language instructions. *arXiv preprint arXiv:2304.14293* (2023).

[197] Frederick J Zimmerman, Dimitri A Christakis, and Andrew N Meltzoff. 2007. Associations between media viewing and language development in children under age 2 years. *The Journal of pediatrics* 151, 4 (2007), 364–368.

A APPENDIX: STORY GENERATION PROMPTS

We provide the complete set of story generation prompts that we used for OSOS. For each prompt, the following stylizations are applied to annotate the prompts and responses for easier understanding.

- underlined text: placeholders to be replaced with the given input text.
- yellow highlighted text : major difference incurred by the prompt.
- **bold-faced words**: the target words given by the Extractor (§5.2).



Appendix

GPT prompt issued by Story Steering Console upon "Extract storybook abstract"

```
You are going to summarize a Korean story in English with 3 sentences. You are going to ignore unique details from the
story such as names. Instead of representing a character as a child, choose either boy or girl. The story is :
(Existing story in Korean)
```

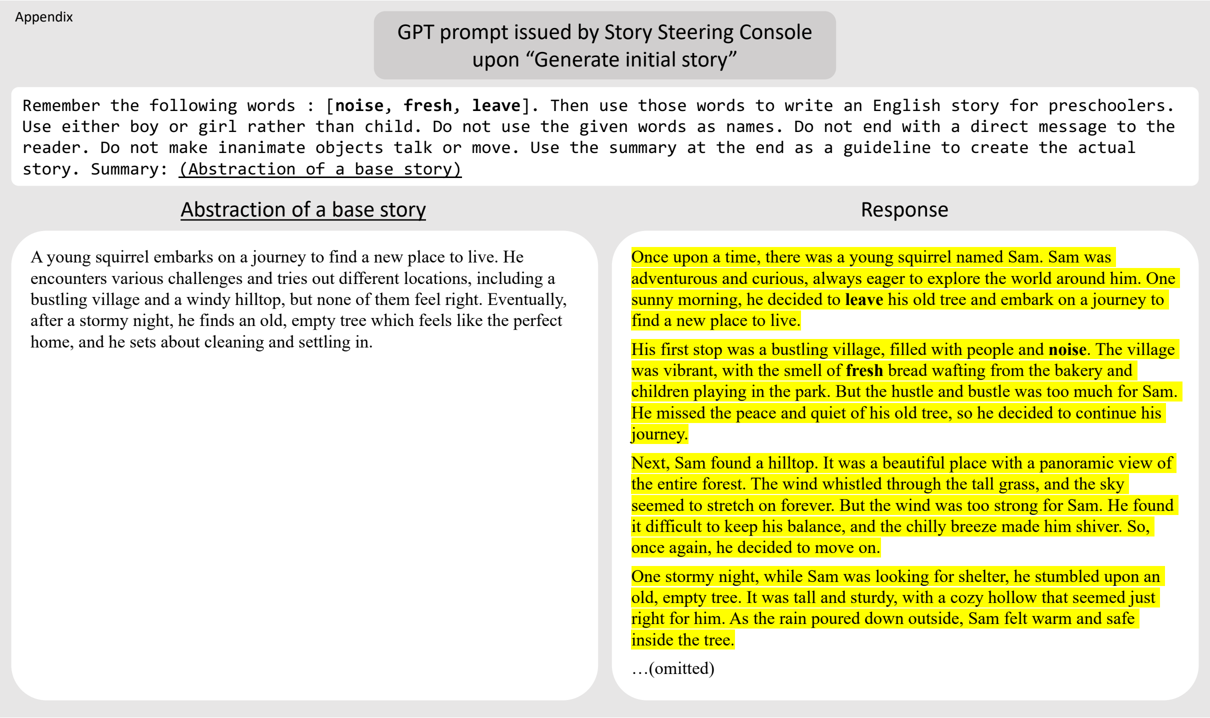
Existing story in Korean

(Existing story in Korean)

Response

A young squirrel embarks on a journey to find a new place to live. He encounters various challenges and tries out different locations, including a bustling village and a windy hilltop, but none of them feel right. Eventually, after a stormy night, he finds an old, empty tree which feels like the perfect home, and he sets about cleaning and settling in.

Figure 18: Story generation prompt auto-issued by Story Steering Console, upon clicking "Extract storybook abstract". This prompt generates a storybook abstract extracted from the given existing story.



Appendix

GPT prompt issued by Story Steering Console upon "Generate initial story"

```
Remember the following words : [noise, fresh, leave]. Then use those words to write an English story for preschoolers.
Use either boy or girl rather than child. Do not use the given words as names. Do not end with a direct message to the
reader. Do not make inanimate objects talk or move. Use the summary at the end as a guideline to create the actual
story. Summary: (Abstraction of a base story)
```

Abstraction of a base story

A young squirrel embarks on a journey to find a new place to live. He encounters various challenges and tries out different locations, including a bustling village and a windy hilltop, but none of them feel right. Eventually, after a stormy night, he finds an old, empty tree which feels like the perfect home, and he sets about cleaning and settling in.

Response

Once upon a time, there was a young squirrel named Sam. Sam was adventurous and curious, always eager to explore the world around him. One sunny morning, he decided to **leave** his old tree and embark on a journey to find a new place to live.

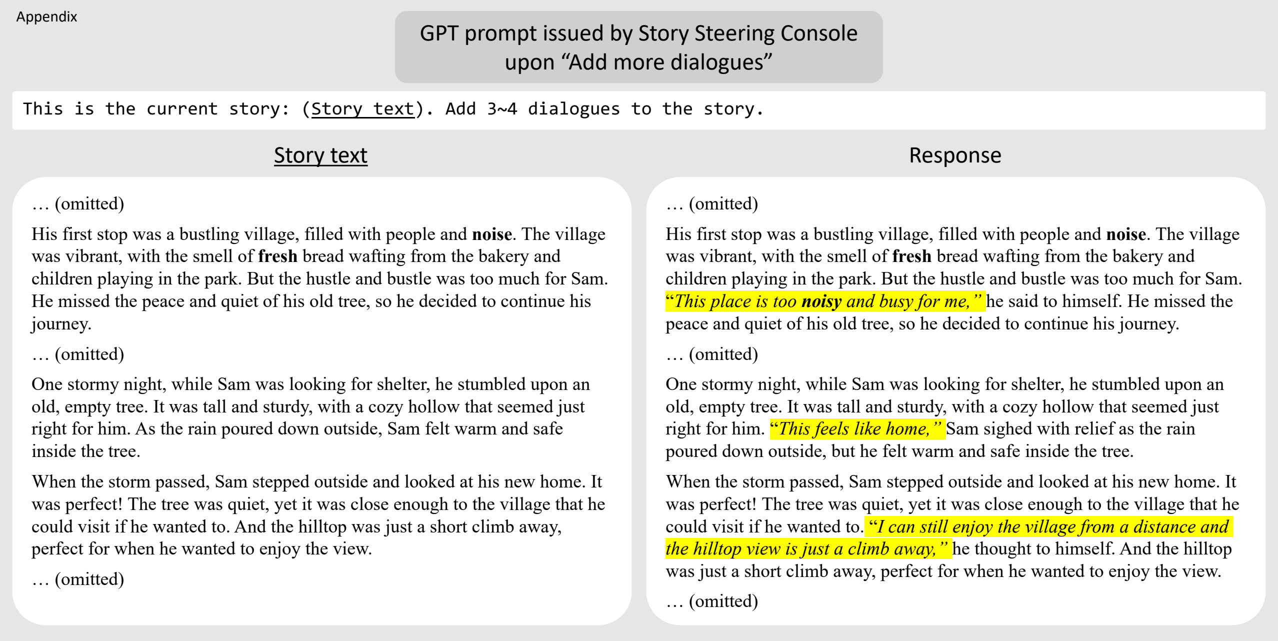
His first stop was a bustling village, filled with people and **noise**. The village was vibrant, with the smell of **fresh** bread wafting from the bakery and children playing in the park. But the hustle and bustle was too much for Sam. He missed the peace and quiet of his old tree, so he decided to continue his journey.

Next, Sam found a hilltop. It was a beautiful place with a panoramic view of the entire forest. The wind whistled through the tall grass, and the sky seemed to stretch on forever. But the wind was too strong for Sam. He found it difficult to keep his balance, and the chilly breeze made him shiver. So, once again, he decided to move on.

One stormy night, while Sam was looking for shelter, he stumbled upon an old, empty tree. It was tall and sturdy, with a cozy hollow that seemed just right for him. As the rain poured down outside, Sam felt warm and safe inside the tree.

…(omitted)

Figure 19: Story generation prompt auto-issued by Story Steering Console, upon clicking "Generate initial story". This prompt generates the first draft of the story from the given storybook abstract.



Appendix

GPT prompt issued by Story Steering Console
upon "Add more dialogues"

`This is the current story: (Story text). Add 3~4 dialogues to the story.`

Story text

… (omitted)

His first stop was a bustling village, filled with people and **noise**. The village was vibrant, with the smell of **fresh** bread wafting from the bakery and children playing in the park. But the hustle and bustle was too much for Sam. He missed the peace and quiet of his old tree, so he decided to continue his journey.

… (omitted)

One stormy night, while Sam was looking for shelter, he stumbled upon an old, empty tree. It was tall and sturdy, with a cozy hollow that seemed just right for him. As the rain poured down outside, Sam felt warm and safe inside the tree.

When the storm passed, Sam stepped outside and looked at his new home. It was perfect! The tree was quiet, yet it was close enough to the village that he could visit if he wanted to. And the hilltop was just a short climb away, perfect for when he wanted to enjoy the view.

… (omitted)

Response

… (omitted)

His first stop was a bustling village, filled with people and **noise**. The village was vibrant, with the smell of **fresh** bread wafting from the bakery and children playing in the park. But the hustle and bustle was too much for Sam. "*This place is too **noisy** and busy for me,"* he said to himself. He missed the peace and quiet of his old tree, so he decided to continue his journey.

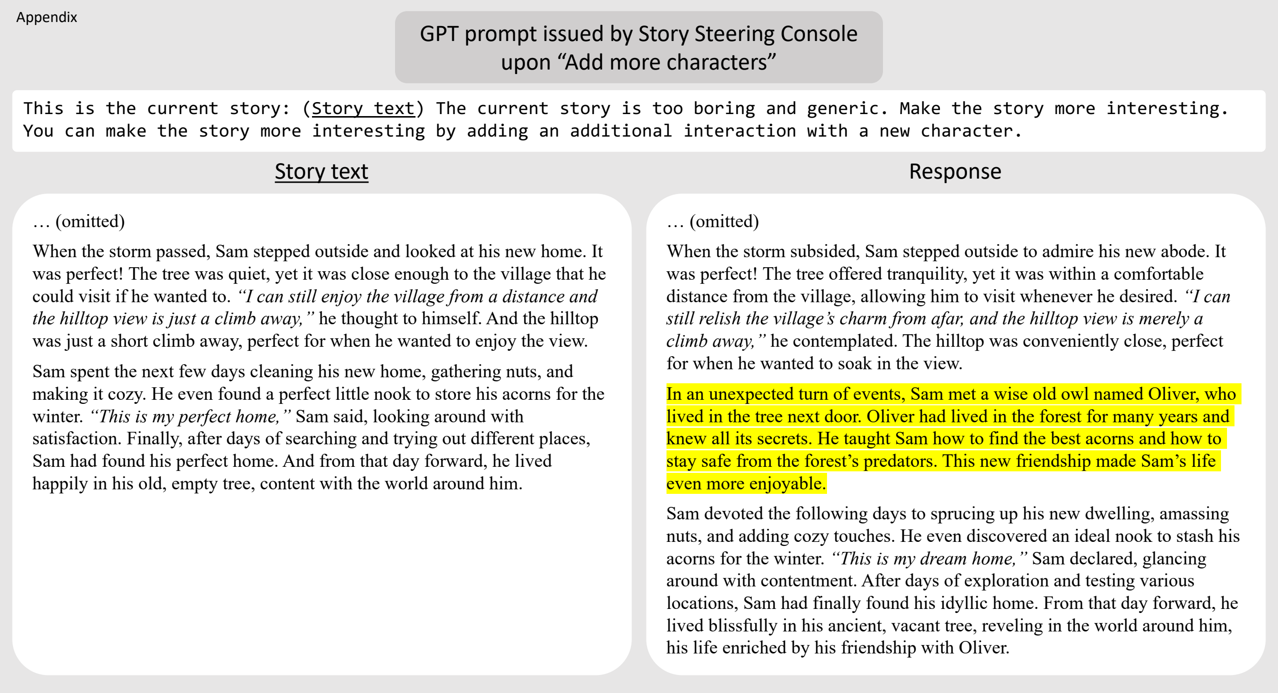
… (omitted)

One stormy night, while Sam was looking for shelter, he stumbled upon an old, empty tree. It was tall and sturdy, with a cozy hollow that seemed just right for him. *"This feels like home,"* Sam sighed with relief as the rain poured down outside, but he felt warm and safe inside the tree.

When the storm passed, Sam stepped outside and looked at his new home. It was perfect! The tree was quiet, yet it was close enough to the village that he could visit if he wanted to. *"I can still enjoy the village from a distance and the hilltop view is just a climb away,"* he thought to himself. And the hilltop was just a short climb away, perfect for when he wanted to enjoy the view.

… (omitted)

Figure 20: Story generation prompt auto-issued by Story Steering Console, upon clicking "Add more dialogues". This prompt adds more dialogues to the given story.



Appendix

GPT prompt issued by Story Steering Console
upon "Add more characters"

```
This is the current story: (Story text) The current story is too boring and generic. Make the story more interesting.
You can make the story more interesting by adding an additional interaction with a new character.
```

Story text

… (omitted)

When the storm passed, Sam stepped outside and looked at his new home. It was perfect! The tree was quiet, yet it was close enough to the village that he could visit if he wanted to. *"I can still enjoy the village from a distance and the hilltop view is just a climb away,"* he thought to himself. And the hilltop was just a short climb away, perfect for when he wanted to enjoy the view.

Sam spent the next few days cleaning his new home, gathering nuts, and making it cozy. He even found a perfect little nook to store his acorns for the winter. *"This is my perfect home,"* Sam said, looking around with satisfaction. Finally, after days of searching and trying out different places, Sam had found his perfect home. And from that day forward, he lived happily in his old, empty tree, content with the world around him.

Response

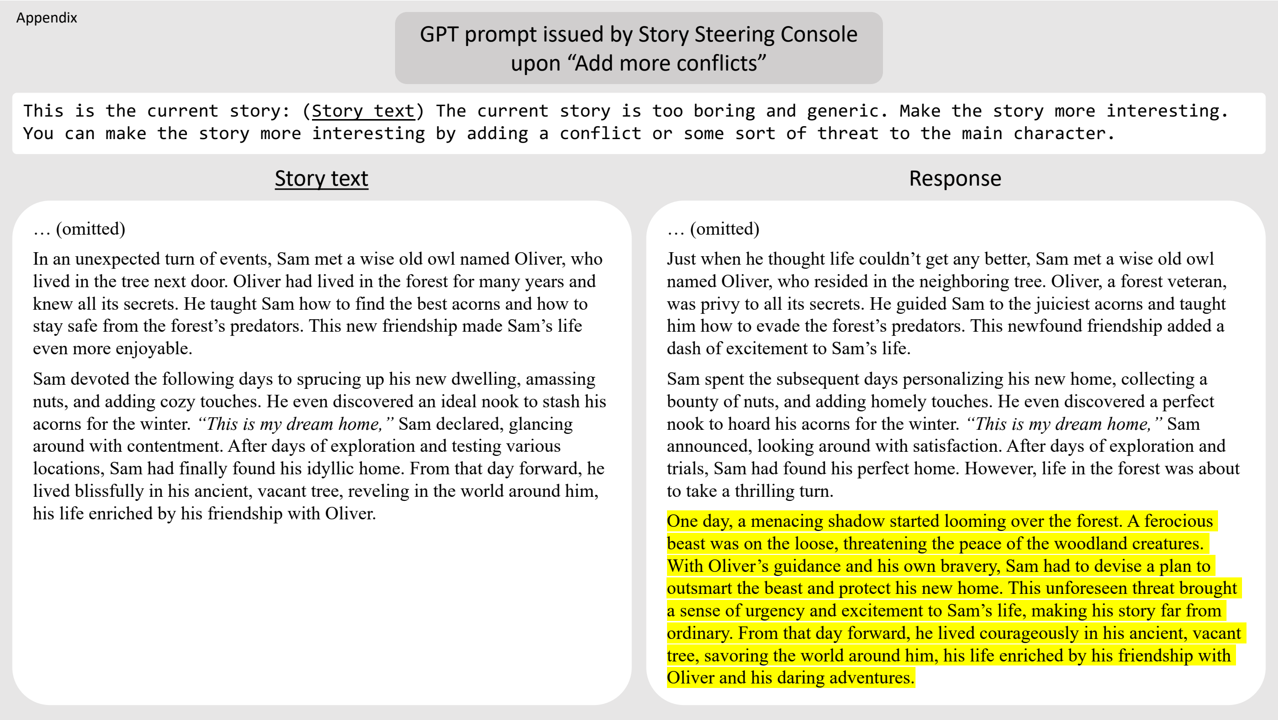
… (omitted)

When the storm subsided, Sam stepped outside to admire his new abode. It was perfect! The tree offered tranquility, yet it was within a comfortable distance from the village, allowing him to visit whenever he desired. *"I can still relish the village's charm from afar, and the hilltop view is merely a climb away,"* he contemplated. The hilltop was conveniently close, perfect for when he wanted to soak in the view.

In an unexpected turn of events, Sam met a wise old owl named Oliver, who lived in the tree next door. Oliver had lived in the forest for many years and knew all its secrets. He taught Sam how to find the best acorns and how to stay safe from the forest's predators. This new friendship made Sam's life even more enjoyable.

Sam devoted the following days to sprucing up his new dwelling, amassing nuts, and adding cozy touches. He even discovered an ideal nook to stash his acorns for the winter. *"This is my dream home,"* Sam declared, glancing around with contentment. After days of exploration and testing various locations, Sam had finally found his idyllic home. From that day forward, he lived blissfully in his ancient, vacant tree, reveling in the world around him, his life enriched by his friendship with Oliver.

Figure 21: Story generation prompt auto-issued by Story Steering Console, upon clicking "Add more characters". This prompt adds more characters to the given story.



Appendix

GPT prompt issued by Story Steering Console upon “Add more conflicts”

`This is the current story: (Story text) The current story is too boring and generic. Make the story more interesting. You can make the story more interesting by adding a conflict or some sort of threat to the main character.`

Story text

… (omitted)

In an unexpected turn of events, Sam met a wise old owl named Oliver, who lived in the tree next door. Oliver had lived in the forest for many years and knew all its secrets. He taught Sam how to find the best acorns and how to stay safe from the forest’s predators. This new friendship made Sam’s life even more enjoyable.

Sam devoted the following days to sprucing up his new dwelling, amassing nuts, and adding cozy touches. He even discovered an ideal nook to stash his acorns for the winter. *“This is my dream home,”* Sam declared, glancing around with contentment. After days of exploration and testing various locations, Sam had finally found his idyllic home. From that day forward, he lived blissfully in his ancient, vacant tree, reveling in the world around him, his life enriched by his friendship with Oliver.

Response

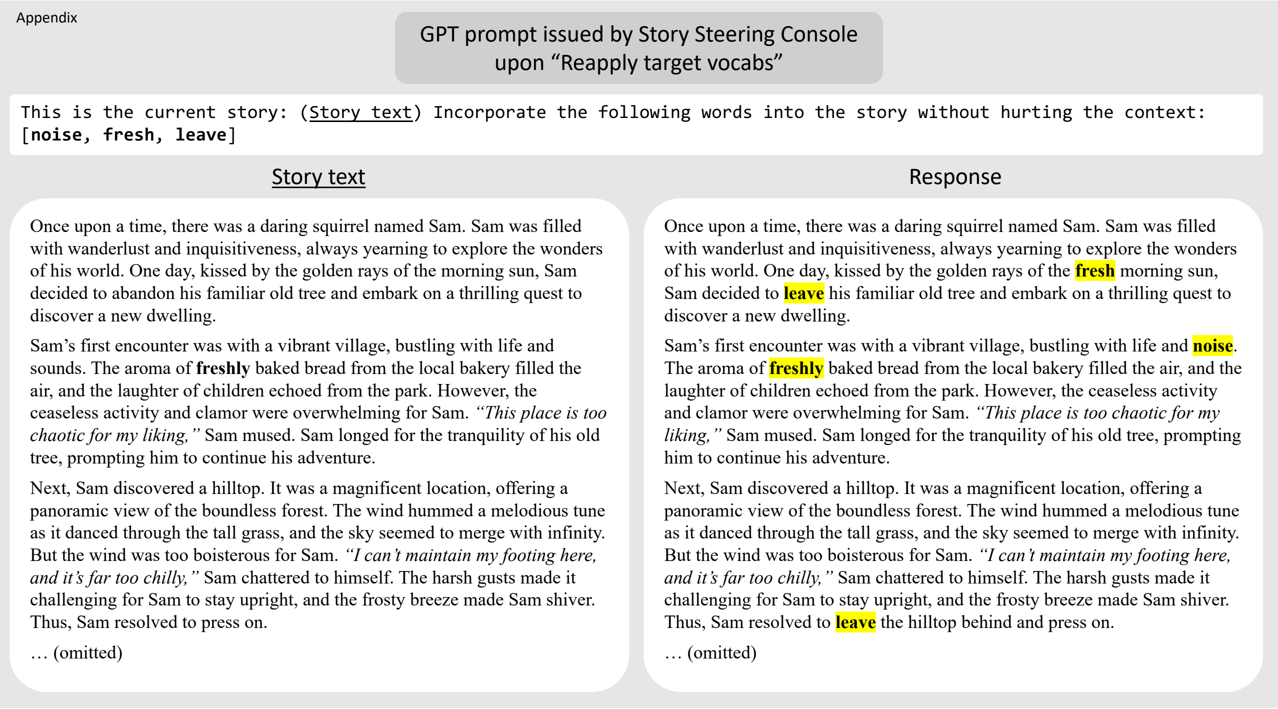
… (omitted)

Just when he thought life couldn’t get any better, Sam met a wise old owl named Oliver, who resided in the neighboring tree. Oliver, a forest veteran, was privy to all its secrets. He guided Sam to the juiciest acorns and taught him how to evade the forest’s predators. This newfound friendship added a dash of excitement to Sam’s life.

Sam spent the subsequent days personalizing his new home, collecting a bounty of nuts, and adding homely touches. He even discovered a perfect nook to hoard his acorns for the winter. *“This is my dream home,”* Sam announced, looking around with satisfaction. After days of exploration and trials, Sam had found his perfect home. However, life in the forest was about to take a thrilling turn.

One day, a menacing shadow started looming over the forest. A ferocious beast was on the loose, threatening the peace of the woodland creatures. With Oliver’s guidance and his own bravery, Sam had to devise a plan to outsmart the beast and protect his new home. This unforeseen threat brought a sense of urgency and excitement to Sam’s life, making his story far from ordinary. From that day forward, he lived courageously in his ancient, vacant tree, savoring the world around him, his life enriched by his friendship with Oliver and his daring adventures.

Figure 22: Story generation prompt auto-issued by Story Steering Console, upon clicking “Add more conflicts”. This prompt adds more conflicts to the given story.



Appendix

GPT prompt issued by Story Steering Console upon “Reapply target vocabs”

`This is the current story: (Story text) Incorporate the following words into the story without hurting the context: [noise, fresh, leave]`

Story text

Once upon a time, there was a daring squirrel named Sam. Sam was filled with wanderlust and inquisitiveness, always yearning to explore the wonders of his world. One day, kissed by the golden rays of the morning sun, Sam decided to abandon his familiar old tree and embark on a thrilling quest to discover a new dwelling.

Sam’s first encounter was with a vibrant village, bustling with life and sounds. The aroma of **freshly** baked bread from the local bakery filled the air, and the laughter of children echoed from the park. However, the ceaseless activity and clamor were overwhelming for Sam. *“This place is too chaotic for my liking,”* Sam mused. Sam longed for the tranquility of his old tree, prompting him to continue his adventure.

Next, Sam discovered a hilltop. It was a magnificent location, offering a panoramic view of the boundless forest. The wind hummed a melodious tune as it danced through the tall grass, and the sky seemed to merge with infinity. But the wind was too boisterous for Sam. *“I can’t maintain my footing here, and it’s far too chilly,”* Sam chattered to himself. The harsh gusts made it challenging for Sam to stay upright, and the frosty breeze made Sam shiver. Thus, Sam resolved to press on.

… (omitted)

Response

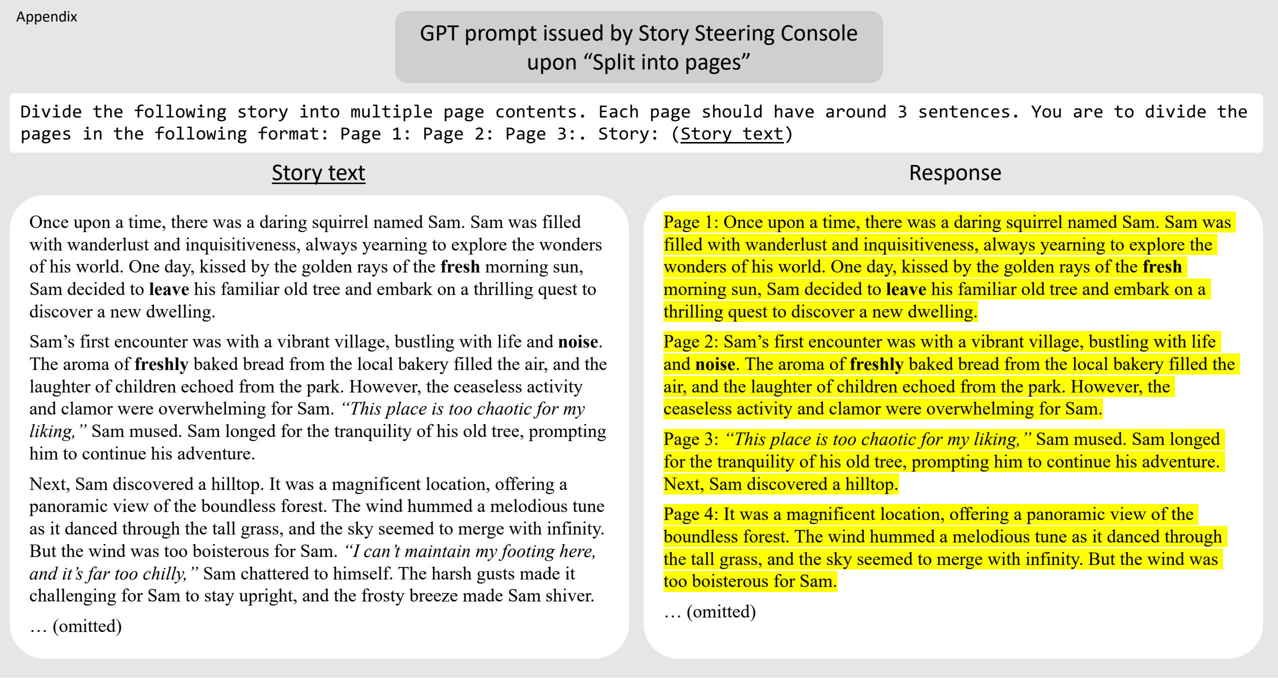
Once upon a time, there was a daring squirrel named Sam. Sam was filled with wanderlust and inquisitiveness, always yearning to explore the wonders of his world. One day, kissed by the golden rays of the **fresh** morning sun, Sam decided to **leave** his familiar old tree and embark on a thrilling quest to discover a new dwelling.

Sam’s first encounter was with a vibrant village, bustling with life and **noise**. The aroma of **freshly** baked bread from the local bakery filled the air, and the laughter of children echoed from the park. However, the ceaseless activity and clamor were overwhelming for Sam. *“This place is too chaotic for my liking,”* Sam mused. Sam longed for the tranquility of his old tree, prompting him to continue his adventure.

Next, Sam discovered a hilltop. It was a magnificent location, offering a panoramic view of the boundless forest. The wind hummed a melodious tune as it danced through the tall grass, and the sky seemed to merge with infinity. But the wind was too boisterous for Sam. *“I can’t maintain my footing here, and it’s far too chilly,”* Sam chattered to himself. The harsh gusts made it challenging for Sam to stay upright, and the frosty breeze made Sam shiver. Thus, Sam resolved to **leave** the hilltop behind and press on.

… (omitted)

Figure 23: Story generation prompt auto-issued by Story Steering Console, upon clicking “Reapply target vocabs”. This prompt reapplies target words to the given story.



Appendix

GPT prompt issued by Story Steering Console
upon "Split into pages"

```
Divide the following story into multiple page contents. Each page should have around 3 sentences. You are to divide the
pages in the following format: Page 1: Page 2: Page 3:. Story: (Story text)
```

Story text

Once upon a time, there was a daring squirrel named Sam. Sam was filled with wanderlust and inquisitiveness, always yearning to explore the wonders of his world. One day, kissed by the golden rays of the **fresh** morning sun, Sam decided to **leave** his familiar old tree and embark on a thrilling quest to discover a new dwelling.

Sam's first encounter was with a vibrant village, bustling with life and **noise**. The aroma of **freshly** baked bread from the local bakery filled the air, and the laughter of children echoed from the park. However, the ceaseless activity and clamor were overwhelming for Sam. *"This place is too chaotic for my liking,"* Sam mused. Sam longed for the tranquility of his old tree, prompting him to continue his adventure.

Next, Sam discovered a hilltop. It was a magnificent location, offering a panoramic view of the boundless forest. The wind hummed a melodious tune as it danced through the tall grass, and the sky seemed to merge with infinity. But the wind was too boisterous for Sam. *"I can't maintain my footing here, and it's far too chilly,"* Sam chattered to himself. The harsh gusts made it challenging for Sam to stay upright, and the frosty breeze made Sam shiver.

… (omitted)

Response

Page 1: Once upon a time, there was a daring squirrel named Sam. Sam was filled with wanderlust and inquisitiveness, always yearning to explore the wonders of his world. One day, kissed by the golden rays of the **fresh** morning sun, Sam decided to **leave** his familiar old tree and embark on a thrilling quest to discover a new dwelling.

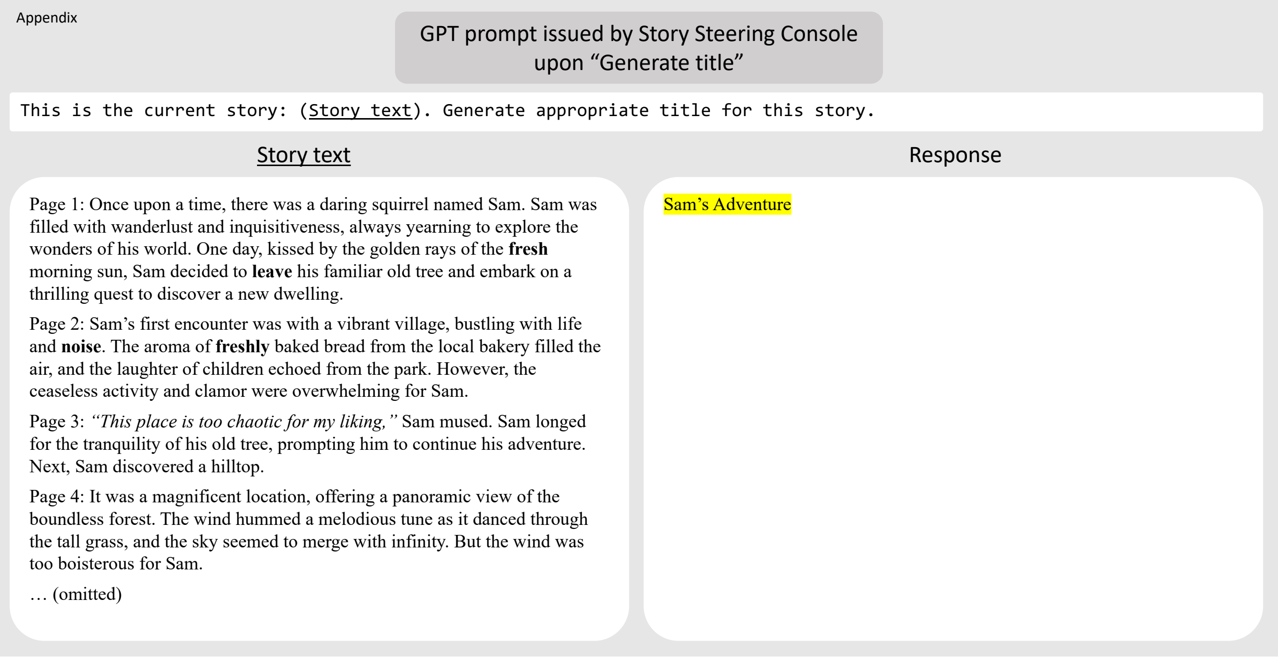
Page 2: Sam's first encounter was with a vibrant village, bustling with life and **noise**. The aroma of **freshly** baked bread from the local bakery filled the air, and the laughter of children echoed from the park. However, the ceaseless activity and clamor were overwhelming for Sam.

Page 3: *"This place is too chaotic for my liking,"* Sam mused. Sam longed for the tranquility of his old tree, prompting him to continue his adventure. Next, Sam discovered a hilltop.

Page 4: It was a magnificent location, offering a panoramic view of the boundless forest. The wind hummed a melodious tune as it danced through the tall grass, and the sky seemed to merge with infinity. But the wind was too boisterous for Sam.

… (omitted)

Figure 24: Story generation prompt auto-issued by Story Steering Console, upon clicking "Split into pages". This prompt paginates the given story, and formats the output in a specific manner: each paginated block starts with a page identifier (e.g., Page 1).



Appendix

GPT prompt issued by Story Steering Console
upon "Generate title"

```
This is the current story: (Story text). Generate appropriate title for this story.
```

Story text

Page 1: Once upon a time, there was a daring squirrel named Sam. Sam was filled with wanderlust and inquisitiveness, always yearning to explore the wonders of his world. One day, kissed by the golden rays of the **fresh** morning sun, Sam decided to **leave** his familiar old tree and embark on a thrilling quest to discover a new dwelling.

Page 2: Sam's first encounter was with a vibrant village, bustling with life and **noise**. The aroma of **freshly** baked bread from the local bakery filled the air, and the laughter of children echoed from the park. However, the ceaseless activity and clamor were overwhelming for Sam.

Page 3: *"This place is too chaotic for my liking,"* Sam mused. Sam longed for the tranquility of his old tree, prompting him to continue his adventure. Next, Sam discovered a hilltop.

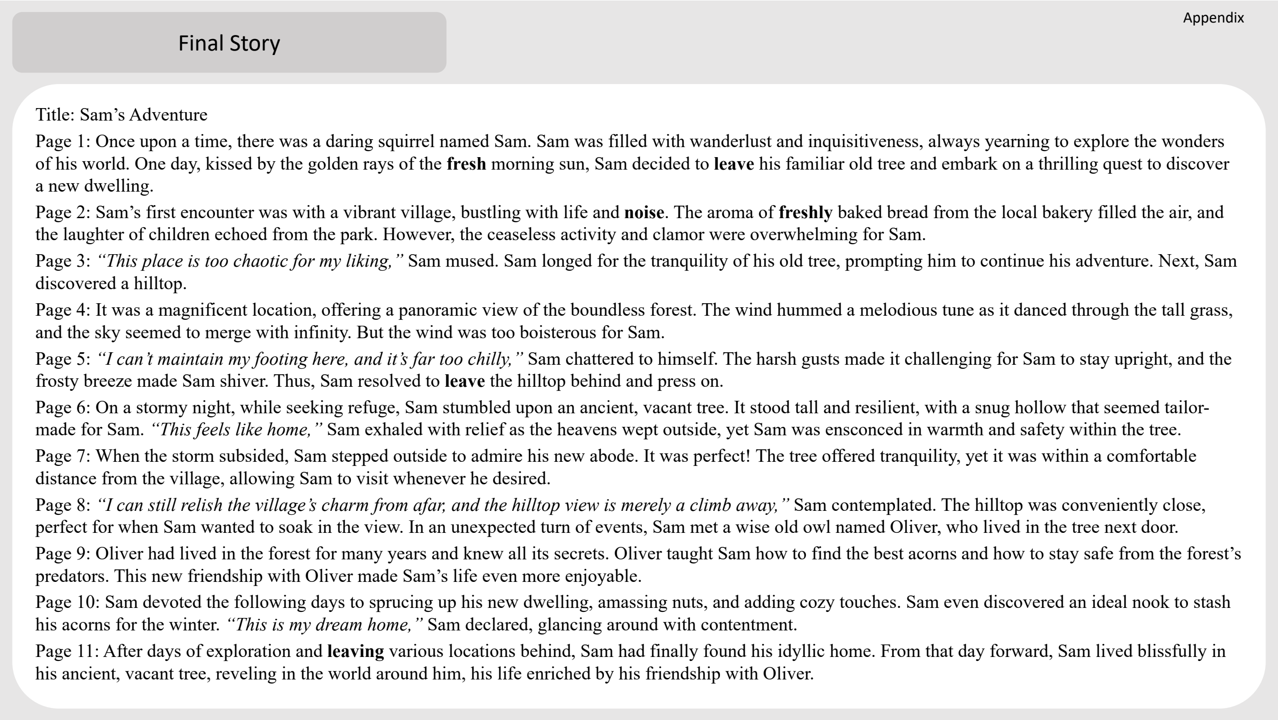
Page 4: It was a magnificent location, offering a panoramic view of the boundless forest. The wind hummed a melodious tune as it danced through the tall grass, and the sky seemed to merge with infinity. But the wind was too boisterous for Sam.

… (omitted)

Response

Sam's Adventure

Figure 25: Story generation prompt auto-issued by Story Steering Console, upon clicking "Generate title". This prompt generates a title for the given story.


Appendix
Final Story

Title: Sam's Adventure

Page 1: Once upon a time, there was a daring squirrel named Sam. Sam was filled with wanderlust and inquisitiveness, always yearning to explore the wonders of his world. One day, kissed by the golden rays of the **fresh** morning sun, Sam decided to **leave** his familiar old tree and embark on a thrilling quest to discover a new dwelling.

Page 2: Sam's first encounter was with a vibrant village, bustling with life and **noise**. The aroma of **freshly** baked bread from the local bakery filled the air, and the laughter of children echoed from the park. However, the ceaseless activity and clamor were overwhelming for Sam.

Page 3: *"This place is too chaotic for my liking,"* Sam mused. Sam longed for the tranquility of his old tree, prompting him to continue his adventure. Next, Sam discovered a hilltop.

Page 4: It was a magnificent location, offering a panoramic view of the boundless forest. The wind hummed a melodious tune as it danced through the tall grass, and the sky seemed to merge with infinity. But the wind was too boisterous for Sam.

Page 5: *"I can't maintain my footing here, and it's far too chilly,"* Sam chattered to himself. The harsh gusts made it challenging for Sam to stay upright, and the frosty breeze made Sam shiver. Thus, Sam resolved to **leave** the hilltop behind and press on.

Page 6: On a stormy night, while seeking refuge, Sam stumbled upon an ancient, vacant tree. It stood tall and resilient, with a snug hollow that seemed tailor-made for Sam. *"This feels like home,"* Sam exhaled with relief as the heavens wept outside, yet Sam was ensconced in warmth and safety within the tree.

Page 7: When the storm subsided, Sam stepped outside to admire his new abode. It was perfect! The tree offered tranquility, yet it was within a comfortable distance from the village, allowing Sam to visit whenever he desired.

Page 8: *"I can still relish the village's charm from afar, and the hilltop view is merely a climb away,"* Sam contemplated. The hilltop was conveniently close, perfect for when Sam wanted to soak in the view. In an unexpected turn of events, Sam met a wise old owl named Oliver, who lived in the tree next door.

Page 9: Oliver had lived in the forest for many years and knew all its secrets. Oliver taught Sam how to find the best acorns and how to stay safe from the forest's predators. This new friendship with Oliver made Sam's life even more enjoyable.

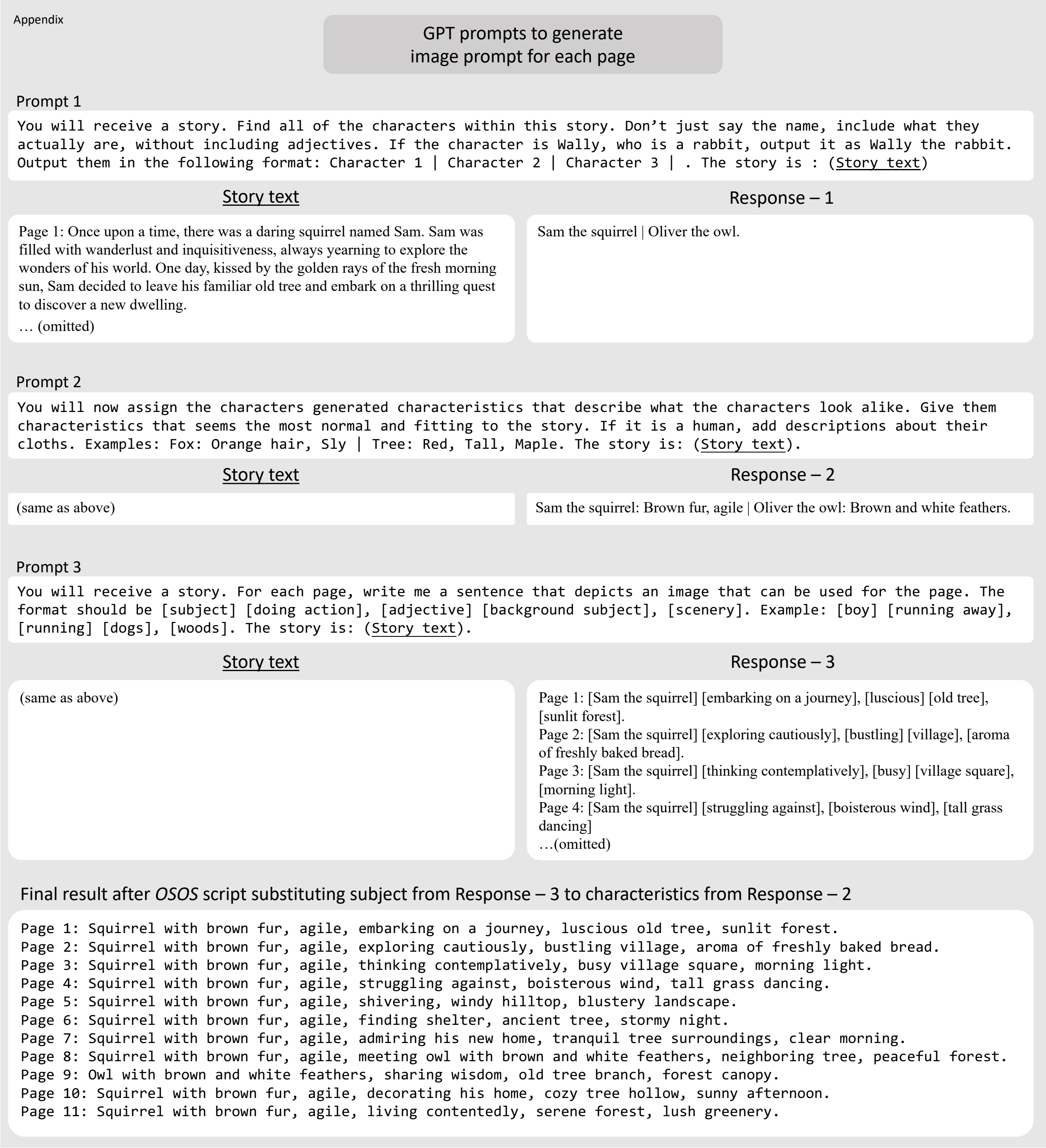
Page 10: Sam devoted the following days to sprucing up his new dwelling, amassing nuts, and adding cozy touches. Sam even discovered an ideal nook to stash his acorns for the winter. *"This is my dream home,"* Sam declared, glancing around with contentment.

Page 11: After days of exploration and **leaving** various locations behind, Sam had finally found his idyllic home. From that day forward, Sam lived blissfully in his ancient, vacant tree, reveling in the world around him, his life enriched by his friendship with Oliver.

Figure 26: Final story generated.

B APPENDIX: IMAGE GENERATION PROMPTS

B.1 Generating Image Prompts



Appendix

GPT prompts to generate
image prompt for each page

Prompt 1

`You will receive a story. Find all of the characters within this story. Don't just say the name, include what they actually are, without including adjectives. If the character is Wally, who is a rabbit, output it as Wally the rabbit. Output them in the following format: Character 1 | Character 2 | Character 3 | . The story is : (Story text)`

Story text	Response – 1
Page 1: Once upon a time, there was a daring squirrel named Sam. Sam was filled with wanderlust and inquisitiveness, always yearning to explore the wonders of his world. One day, kissed by the golden rays of the fresh morning sun, Sam decided to leave his familiar old tree and embark on a thrilling quest to discover a new dwelling. … (omitted)	Sam the squirrel | Oliver the owl.

Prompt 2

`You will now assign the characters generated characteristics that describe what the characters look alike. Give them characteristics that seems the most normal and fitting to the story. If it is a human, add descriptions about their cloths. Examples: Fox: Orange hair, Sly | Tree: Red, Tall, Maple. The story is: (Story text).`

Story text	Response – 2
(same as above)	Sam the squirrel: Brown fur, agile | Oliver the owl: Brown and white feathers.

Prompt 3

`You will receive a story. For each page, write me a sentence that depicts an image that can be used for the page. The format should be [subject] [doing action], [adjective] [background subject], [scenery]. Example: [boy] [running away], [running] [dogs], [woods]. The story is: (Story text).`

Story text	Response – 3
(same as above)	Page 1: [Sam the squirrel] [embarking on a journey], [luscious] [old tree], [sunlit forest]. Page 2: [Sam the squirrel] [exploring cautiously], [bustling] [village], [aroma of freshly baked bread]. Page 3: [Sam the squirrel] [thinking contemplatively], [busy] [village square], [morning light]. Page 4: [Sam the squirrel] [struggling against], [boisterous wind], [tall grass dancing] …(omitted)

Final result after *OSOS* script substituting subject from Response – 3 to characteristics from Response – 2

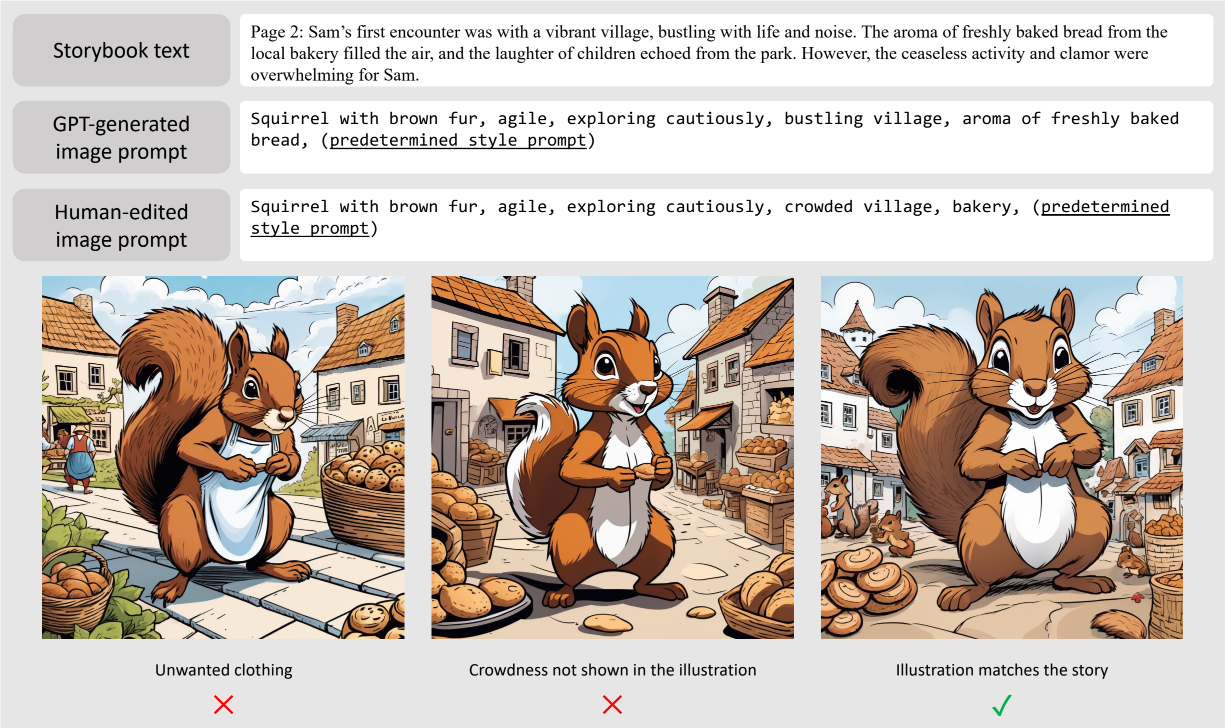
```
Page 1: Squirrel with brown fur, agile, embarking on a journey, luscious old tree, sunlit forest.
Page 2: Squirrel with brown fur, agile, exploring cautiously, bustling village, aroma of freshly baked bread.
Page 3: Squirrel with brown fur, agile, thinking contemplatively, busy village square, morning light.
Page 4: Squirrel with brown fur, agile, struggling against, boisterous wind, tall grass dancing.
Page 5: Squirrel with brown fur, agile, shivering, windy hilltop, blustery landscape.
Page 6: Squirrel with brown fur, agile, finding shelter, ancient tree, stormy night.
Page 7: Squirrel with brown fur, agile, admiring his new home, tranquil tree surroundings, clear morning.
Page 8: Squirrel with brown fur, agile, meeting owl with brown and white feathers, neighboring tree, peaceful forest.
Page 9: Owl with brown and white feathers, sharing wisdom, old tree branch, forest canopy.
Page 10: Squirrel with brown fur, agile, decorating his home, cozy tree hollow, sunny afternoon.
Page 11: Squirrel with brown fur, agile, living contentedly, serene forest, lush greenery.
```

Figure 27: A series of GPT prompts that generate an image prompt for each page. In Prompt 1, each character is given an annotation of what they are. In Prompt 2, each annotated character is given a few visual descriptions inferred by GPT. In Prompt 3, a formatted visual description is generated for each paginated text. The final result is obtained by appending the visual descriptions from Prompt 2 to the character tokens in Prompt 3.

B.2 Generating Images

We curate the image-generation prompts for select pages. Each generated image prompt has "(predetermined style prompt)", which is shared across all images of a book. This style prompt is chosen only once for a newly generated story by the human-in-the-loop. The human chooses a style out of preset options (e.g., Comic book style, Photographic style, etc.) deemed most suitable with the generated storybook text. Having this style prompt appended to all the subsequent per-page image-generation prompts helps the images drawn in a consistent illustration style throughout the entire storybook.

A predetermined style prompt can be easily and quickly added through our interface. Once added, the interface automatically converts it into the corresponding preset of style properties containing both *positive prompt* and *negative prompt*, and appends them to the image prompt. The *positive prompt* directs the Stable Diffusion of the desired outcome, whereas the *negative prompt* steers it away from undesired results. In the following samples, the predetermined style prompt is "Comic book style", which makes the interface automatically add "comic, graphic illustration, comic art, graphic novel art, vibrant, highly detailed" as a *positive prompt*, and "photograph, deformed, glitch, noisy, realistic, stock photo" as a *negative prompt*.



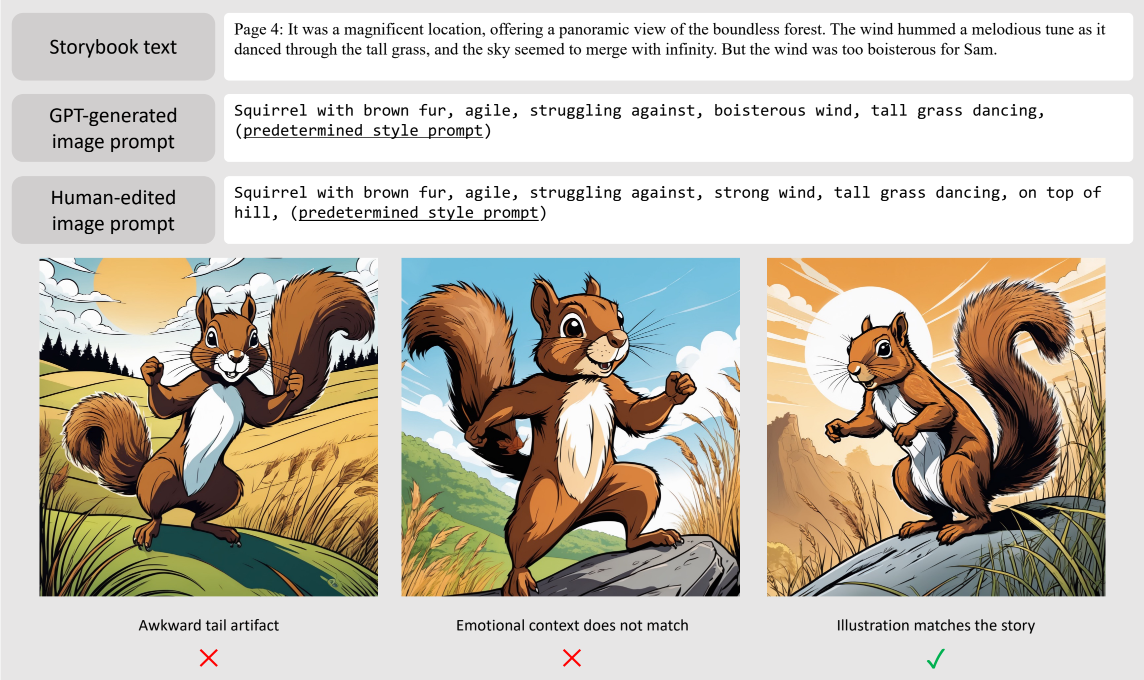
Storybook text	Page 2: Sam's first encounter was with a vibrant village, bustling with life and noise. The aroma of freshly baked bread from the local bakery filled the air, and the laughter of children echoed from the park. However, the ceaseless activity and clamor were overwhelming for Sam.
GPT-generated image prompt	`Squirrel with brown fur, agile, exploring cautiously, bustling village, aroma of freshly baked bread, (predetermined style prompt)`
Human-edited image prompt	`Squirrel with brown fur, agile, exploring cautiously, crowded village, bakery, (predetermined style prompt)`

Unwanted clothing ✗

Crowdness not shown in the illustration ✗

Illustration matches the story ✓

Figure 28: Image generation prompt to create an illustration for page 2. The GPT-generated image prompts are obtained by the process shown in Figure 27. Once chosen by the human reviewer, the pre-determined style prompt is auto-appended throughout all subsequent pages. The human may apply minor edits to the GPT-generated image prompts. Finally, a batch of images is generated by Stable Diffusion, as shown above. The human reviewer picks the most appropriate one for this page.



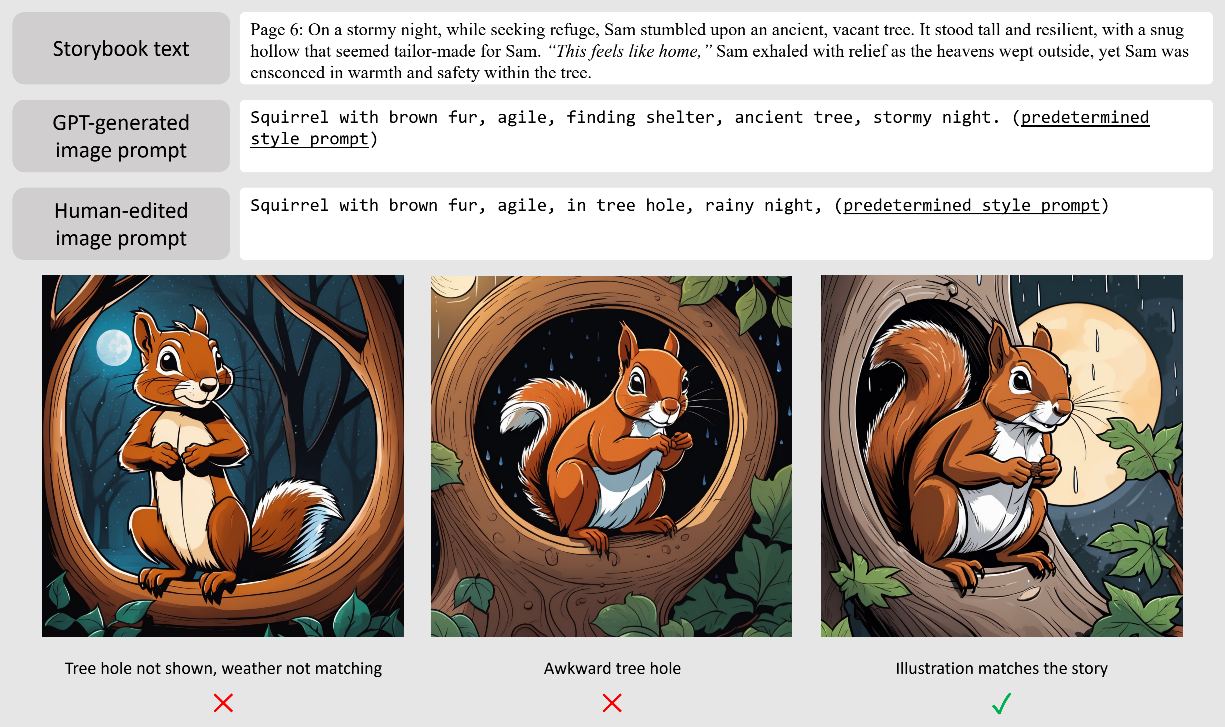
Storybook text	Page 4: It was a magnificent location, offering a panoramic view of the boundless forest. The wind hummed a melodious tune as it danced through the tall grass, and the sky seemed to merge with infinity. But the wind was too boisterous for Sam.
GPT-generated image prompt	`Squirrel with brown fur, agile, struggling against, boisterous wind, tall grass dancing, (predetermined style prompt)`
Human-edited image prompt	`Squirrel with brown fur, agile, struggling against, strong wind, tall grass dancing, on top of hill, (predetermined style prompt)`

Awkward tail artifact ✗

Emotional context does not match ✗

Illustration matches the story ✓

Figure 29: Image generation prompt to create an illustration for page 4. The process is identical to Figure 28, except the pre-determined style prompt which is now auto-appended as the one chosen on the initial page.



Storybook text	Page 6: On a stormy night, while seeking refuge, Sam stumbled upon an ancient, vacant tree. It stood tall and resilient, with a snug hollow that seemed tailor-made for Sam. *"This feels like home,"* Sam exhaled with relief as the heavens wept outside, yet Sam was ensconced in warmth and safety within the tree.
GPT-generated image prompt	`Squirrel with brown fur, agile, finding shelter, ancient tree, stormy night. (predetermined style prompt)`
Human-edited image prompt	`Squirrel with brown fur, agile, in tree hole, rainy night, (predetermined style prompt)`

Tree hole not shown, weather not matching ✗

Awkward tree hole ✗

Illustration matches the story ✓

Figure 30: Image generation prompt to create an illustration for page 6. The process is identical to Figure 28, except the pre-determined style prompt which is now auto-appended as the one chosen on the initial page.