Exploring Automated Keyword Mnemonics Generation with Large Language Models via Overgenerate-and-Rank

Jaewook Lee, Hunter McNichols, Andrew Lan
University of Massachusetts Amherst
{jaewooklee,wmcnichols,andrewlan}@cs.umass.edu

Abstract

In this paper, we study an under-explored area of language and vocabulary learning: keyword mnemonics, a technique for memorizing vocabulary through memorable associations with a target word via a verbal cue. Typically, creating verbal cues requires extensive human effort and is quite time-consuming, necessitating an automated method that is more scalable. We propose a novel overgenerate-and-rank method via prompting large language models (LLMs) to generate verbal cues and then ranking them according to psycholinguistic measures and takeaways from a pilot user study. To assess cue quality, we conduct both an automated evaluation of imageability and coherence, as well as a human evaluation involving English teachers and learners. Results show that LLM-generated mnemonics are comparable to human-generated ones in terms of imageability, coherence, and perceived usefulness, but there remains plenty of room for improvement due to the diversity in background and preference among language learners.

1 Introduction

Recent advances in natural language processing have expanded the exploration of developing automated methods for applications in *language learning*, including second language acquisition (Zhang et al., 2021; Yeung and Lee, 2021; Okano et al., 2023), linguistic skill modeling (Zylich and Lan, 2021), language assessment and correction (Katinskaia and Yangarber, 2021), and language practice through conversational chatbots (Tyen et al., 2022; Liang et al., 2023). In this work, we investigate an intriguing but relatively under-explored vocabulary learning application: keyword mnemonics (KM) (Atkinson and Raugh, 1975).

KM is a widely utilized technique employed by language learners to memorize vocabulary effectively, through the creation of memorable associations with keywords that they already know, aided

Figure 1: An example of human-authored verbal cue (Geer and Geer, 2018) for the target word *"alleviate"* and visual cue generated by Stable Diffusion XL. We study automated verbal cue generation in this work and leave visual cue generation for future work.

by verbal and visual cues. This technique is used in many language learning resources, from traditional books (Burchers et al., 2000; Heisig, 2011; Geer and Geer, 2018) to modern community-based learning platforms (Mnemonic Dictionary, 2024; Koohii Kanji, 2024), where users contribute mnemonics and the community vote on their effectiveness.

KM involves a two-step process that creates an acoustic and imagery link for the learner. For instance, consider a learner learning the target word *"alleviate"*, as shown in Figure 1. First, an acoustic link is established to an already-known set of keywords with similar pronunciation, e.g., *"a, leaf, he, ate."* Second, a imagery link is established to these keywords with a verbal cue that evokes a mental image, e.g., *"On his plate, there was a leaf to alleviate his hunger."* These links can be further reinforced using a visual cue.

Despite its effectiveness, KM generation is challenging since it demands significant human effort to meticulously find keywords that resemble the target word and create memorable verbal cues. For instance, the word *"alleviate"* has a large number of possible keyword combinations, such as *"a,*

5521

leaf, he, ate" or *"a, levy, it"*; sorting through them takes considerable effort from teachers and learners. Moreover, the task of creating verbal cues using the keywords that evoke vivid mental images adds an extra layer of complexity to the process.

Existing work on automated KM generation remains limited in both technique and evaluation. For example, earlier methods are restricted to generating a single keyword in the context of second language acquisition. The work in Savva et al. (2014) uses a cognitive psychology-inspired method to automatically generate a keyword in the first language that has high phonetic, orthographic, and semantic similarity with the target word in the second language. More recently, the emergence of large language models (LLMs) provides a new and potentially more scalable solution to automated KM generation (Lee and Lan, 2023), expanding automated KM generation to generating verbal cues using the keywords generated by Savva et al. (2014). The work in Balepur et al. (2024) fine-tunes LLaMA-2 with verbal cues from an online platform and aligns it using student feedback, highlighting how incorporating student preferences can improve LLM-generated verbal cues.

However, evaluating KMs remains challenging due to two main reasons: the lack of automated metrics to evaluate the quality of verbal cues, and the subjective nature of mnemonics to language learners. We need automated ways to evaluate important aspects of KMs such as imageability and coherence. The work in Wu and Smith (2023) explored using text-to-image models for the automatic assessment of sentence imageability, although it has not been investigated in the context of language learning.

Contributions In this paper, we propose an overgenerate-and-rank method to generate verbal cues for vocabulary learning in first language acquisition (English) using LLMs. For a target word that a learner needs to learn, our method first prompts an LLM to generate a set of syllabic keywords and then generate a corresponding verbal cue. For both keywords and verbal cues, we overgenerate and then rank the candidates, using various ranking criteria that are grounded in both cognitive psychology principles used in previous studies and additional insights gained from a pilot user study. To assess the quality of verbal cues generated by our method, we conduct both 1) an automated evaluation of their imageability and coherence using proxy metrics and 2) a human evaluation with both

English teachers and learners, comparing LLM-generated and human-authored cues on imageability, coherence, and additionally usefulness to language learners. Results indicate that our LLM-generated cues are comparable to (and often better than) human-authored ones. We also conduct several case studies and discuss the varying degrees of human agreement on different aspects of KMs, which highlight the need to further align generated KMs with individual preferences.

2 Problem Statement

We now formally define the verbal cue generation task. To do so, we first need to generate a set of *syllabic keywords* given a target word, then a *verbal cue* that contains them. Given a target word t, we denote its syllables as $\mathcal{S}_t = (s_1, ..., s_L)$ where L is the total number of syllables in t. Our goal is to generate a set of M syllabic keywords as $\mathcal{K} = \{k_1, \ldots, k_M\}$, where k_m is phonetically similar to one (or more) consecutive syllables in t, beginning at index l_m and ending at l'_m, which we denote as $s_{l_m:l'_m}$. The set of syllabic keywords also has to cover all syllables, i.e., $\cup_m \{l_m, \ldots, l'_m\} = \{1, \ldots, L\}$. This task is challenging, since that there are 2^{L-1} total possible ways to split the set of syllables of t, and that we need to choose the right keywords that preserve the phonetic properties of the target word. Then, our task is to generate a verbal cue $\mathcal{V} = (w_1, ..., w_N)$, where w_i denotes the i-th word in the cue. The constraint we put on the verbal cue is that it must contain both the specific target word and all syllabic keywords, which we formally define as

$$\forall x \in \mathcal{K} \cup \{t\}, \exists i \in \{1, \ldots, N\} \text{ s.t. } w_i = x.$$

3 Methodology

We propose a two-step overgenerate-and-rank method for KM generation via LLM prompting to navigate the large space of possible keywords and verbal cues. First, we overgenerate multiple sets of candidate syllabic keywords by prompting an LLM, rank them according to a series of measures, and select the top set. Second, we use these keywords to overgenerate multiple verbal cues, rank them according to insights obtained from a pilot study with English teachers. See Supplementary Material A for the exact prompts we used.

3.1 Keyword Generation

For keyword generation, we craft a prompt that includes the task description, instructing the LLM to generate a set of keywords phonetically similar to the syllables of the target word. The instructions include the following rules: 1) Each keyword must be a complete word, familiar and commonly used at an SAT vocabulary level; 2) The keywords should resemble the target word phonetically when spoken together, even if they don't match the exact number of syllables; 3) The words must not be offensive. We also provide a set of three in-context examples.

We overgenerate up to $2L + 1$ sets of candidate keywords $\hat{\mathcal{K}}$ where L is the number of syllables in the target word. To account for aspects of keyword generation other than phonetic similarity, we also consider the following psychology measures used in prior work (Savva et al., 2014) when ranking keywords: imageability, orthographic similarity, which measures how closely the keywords resemble the spelling of the target word, and semantic similarity, which measures how closely are the meanings of the keywords to that of the target word. For the imageability and semantic similarity rankings, lemmatization is applied to both the keywords and the target words.

To create the imageability ranking R_{img}, we compute the average imageability score (Ljubešić et al., 2018a; Scott et al., 2019) of the keywords, excluding stopwords. We prioritize the Glasgow Norms ratings and adjust the scores from Ljubešić et al. (2018b) to a 7-point scale for consistency. If words are absent from the dataset, we assign a score of 1, representing the lowest value on the scale. The score f_{img} is determined by the sum of imageability scores ($\mathcal{L}_{\text{img}}$) of individual keywords divided by the number of keywords:

$$f_{\text{img}}(\hat{\mathcal{K}}) = 1/|\hat{\mathcal{K}}| \sum_{k \in \mathcal{K}} \mathcal{L}_{\text{img}}[k].$$

A set of keywords is ranked higher in R_{img} if the imageability score is higher; same for the other rankings below unless stated otherwise. To create the orthographic similarity ranking R_{orth}, we calculate the Levenshtein distance between a target word and the concatenated keywords, $D_{\text{lev}}(\cdot, \cdot)$. The score f_{orth} is determined by this distance:

$$f_{\text{orth}}(\hat{\mathcal{K}}, t) = D_{\text{lev}}(\text{concat}(\hat{\mathcal{K}}), t).$$

To create the semantic similarity ranking R_{sem}, we compute the cosine similarity between each keyword's embeddings and the target word's embedding (Bojanowski et al., 2017). The score f_{sem} is determined by this maximum similarity:

$$f_{\text{sem}}(\hat{\mathcal{K}}, t) = \max_{k \in \hat{\mathcal{K}}} \cos(\text{emb}(k), \text{emb}(t)).$$

The final, overall keyword ranking is defined as the geometric mean of the three rankings, i.e., $\sqrt[3]{\left(R_{\text{img}} \cdot R_{\text{orth}} \cdot R_{\text{sem}} \right)}$. We observe empirically that this method works well and combines the rankings without additional parameters to tune. Also, ties in this ranking are rare and we break randomly.

3.2 Pilot Study for Verbal Cue Generation

Since the perceived usefulness of a verbal cue can be highly subjective to each learner and not extensively studied in prior work, we conduct a pilot study with English teachers on both LLM-generated and human-authored verbal cues, to identify important features of verbal cues that they think would help learners. In this pilot test, we use a single in-context example, prompting an LLM for both keywords and a corresponding verbal cue at once. We received four main suggestions that inform our verbal cue generation process:

i. Keep the original keyword order in the cue.

ii. Provide clear context for the target word.

iii. Use words at the same (complexity) level as or lower than the target word.

iv. Keep the cue short; long ones are not helpful.

3.3 Verbal Cue Generation

For verbal cue generation, we craft a prompt with a task description that instructs the LLM to generate a coherent verbal cue containing the target word and all syllabic keywords. In each in-context example, we also include the target word's meaning to avoid homonyms, together with a context explanation for richer contextual information.

We overgenerate up to 5 candidate verbal cues and filter out cases where the syllabic keywords are not in order, according to suggestion i). To account for other suggestions, we rank these candidate verbal cues according to two measures: context completeness and age of acquisition (AoA). The former refers to the extent of contextual information on the target word given by the verbal cue, while the other is a psycholinguistic measure that indicates the typical age at which a word is learned. For both rankings, lemmatization is applied to both the keywords and the target words.

To create the context completeness ranking R_{cont}, we employ a masked modeling technique inspired by suggestion ii). Specifically, we mask out the target word t within a verbal cue $\hat{\mathcal{V}}$ and prompt an LLM to predict the five most likely words under the mask, which are listed in a set $\mathcal{C}$:

$$\mathcal{C} = \text{LLM}_{\text{top-5}}\left(\text{mask}(\hat{\mathcal{V}}, t) \rightarrow t\right).$$

Then, we calculate the average cosine similarity between the word embeddings of each predicted word c and the target word t:

$$f_{\text{cont}}(\mathcal{C}, t) = 1/|\mathcal{C}| \sum_{c \in \mathcal{C}} \cos\left(\text{emb}(c), \text{emb}(t)\right).$$

Intuitively, a high average cosine similarity means that it is easy for the LLM to predict the target word (or similar words) given other words in the verbal cue, which indicates that it contains complete contextual information. To create the AoA ranking R_{AoA}, we sum the AoA ($\mathcal{L}_{\text{AoA}}$) (Kuperman et al., 2012) of words in the verbal cue to establish a ranking, which penalizes complex words in the verbal cue according to suggestion iii). The sum also penalizes long verbal cues according to suggestion iv). We exclude stopwords and disregard words that are not in $\mathcal{L}_{\text{AoA}}$. The word complexity score f_{AoA} is given by:

$$f_{\text{AoA}}(\hat{\mathcal{V}}) = \sum_{w \in \hat{\mathcal{V}}} \mathcal{L}_{\text{AoA}}[w].$$

Unlike all other rankings above, verbal cues with lower AoA scores are ranked higher in R_{AoA}.

The final, overall verbal cue ranking is defined as the geometric mean of the two rankings, i.e., $\sqrt{(R_{\text{cont}} \cdot R_{\text{AoA}})}$.

4 Automated Evaluation

In the following sections, we aim to explore three specific aspects of KMs that can vary between individuals: *imageability*, *coherence*, and *usefulness*. In this section, we introduce proxy metrics for automatically evaluating imageability and coherence. Later, in Section 5, we conduct a human evaluation to additionally measure the perceived usefulness of verbal cues by both English teachers and learners, in addition to imageability and coherence.

4.1 Dataset

Since there is no established baseline for evaluating the three aspects of KMs, we compare with human-authored cues with LLM-generated ones. We utilize the book "Picture These SAT Words!" (Geer and Geer, 2018), which consists of around 300 SAT words and provides keywords, verbal, and visual cues. We randomly sample 60 target words, or one fifth of all words from this book, to use in our experiment, due to the significant cost of human evaluation.

We use GPT-4 (temp= 0.7, top_p= 1) to generate both keywords and verbal cues via overgenerate-and-rank. See Supplementary Material B for all target words used in our evaluation along with LLM-generated and human-authored verbal cues.

4.2 Metrics

We define the evaluation criteria for the three aspects of verbal cues as follows: Imageability assesses the effectiveness of verbal cues in evoking mental images, coherence evaluates the logical consistency of the verbal cues, and usefulness determines how helpful these cues are in aiding a learner to learn the target word.

We employ several automated metrics as proxies for assessing the imageability and coherence of verbal cues. Alongside these metrics, we also check the quality of keywords to compare those generated by LLM with human-authored ones, thereby performing a preliminary validation.

4.2.1 Keywords

We evaluate the quality of keywords by applying three ranking criteria outlined in Section 3.1: **Imageability** (word-level), orthographic similarity (**Orthographic Sim.**), and semantic similarity (**Semantic Sim.**), adopted by prior work (Savva et al., 2014). We also introduce two new criteria: **Syllable Ratio** and phonetic similarity (**Phonetic Sim.**). Syllable ratio is calculated by dividing the number of keywords by the total number of syllables, assessing how well the keywords align with the syllables of the target word. Phonetic similarity, determined using the International Phonetic Alphabet (IPA) (Association, 1999), calculates the Levenshtein distance between the concatenated IPAs of the keywords and that of the target word.

4.2.2 Verbal Cue

For imageability, we adopt a similar methodology to the one introduced in (Wu and Smith, 2023), which generates images from textual sentences using a text-to-image model DALL·E mini (Dayma et al., 2021) and applies the text-image alignment metric CLIP (Radford et al., 2021). In our study, we use a larger text-to-image model, Stable Dif-

Method	Syllable Ratio$^\uparrow$	Phonetic Sim.$^\uparrow$	Imageability$^\uparrow$	Orthographic Sim.$^\uparrow$	Semantic Sim.$^\uparrow$
Barron	0.87	**0.52**	0.51	0.37	0.11
Ours	**0.92**	**0.52**	**0.76**	**0.40**	**0.12**

Table 1: Comparative analysis of syllabic keywords from our method and those in Barron's book (Geer and Geer, 2018), using the three ranking criteria along with two additional metrics: syllable ratio and phonetic similarity. Values are normalized to $[0, 1]$ for ease of comparison.

Method		IMR$^\uparrow$	PPL$^\downarrow$
Keyword	Verbal Cue		
Barron		0.56	444.8
Ours		**0.61**	**156.4**

Table 2: Comparative analysis of verbal cues from our method and Barron's book (Geer and Geer, 2018), using the ImageReward and Perplexity metrics.

fusion 2.0 (Rombach et al., 2021), along with a more advanced text-image alignment metric, ImageReward (Xu et al., 2024). Given a verbal cue $\hat{\mathcal{V}}$, we randomly sample a set of 9 images $\mathcal{I}$ using Stable Diffusion. The imageability of the verbal cue is then set as the maximum ImageReward score (**IMR**) between the verbal cue and the generated images. For coherence, we calculate the perplexity (**PPL**) of the verbal cue using the open-source LLM Llama3-8B (Touvron et al., 2023), since textual coherence correlates with how well a pre-trained LLM can predict a sequence of text tokens.

4.3 Results

In what follows, "Ours" refers to LLM-generated keywords and verbal cues using our overgenerate-and-rank method, whereas "Barron" refers to human-authored cues in Barron's book.

4.3.1 Keywords

Table 1 compares keyword generation performance between our method and humans across all metrics. We see that our keywords are as good or better than human-authored ones across all metrics. In terms of imageability, our method gets a much higher score than human authors, likely because human authors often employs proper nouns (e.g., *"Guy"* for *"beguile"*) or alphabet-based terms (e.g., *"D grade"* for *"degradation"*) that are less imageable. On the other metrics, our method slightly outperforms human authors, with the exception of matching human authors on phonetic similarity. As a case study, for the keyword *"enmity"* (IPA: /'ɛn-mɪti/), the LLM-generated *"hen, mitt, tee"* (IPA: /hɛn/, /mɪt/, /ti/) and the human-authored *"N, mitt, hi"* (IPA: /ɛn/, /mɪt/, /hi/) produces keywords with the same Levenshtein distances in the IPA space from the target word. This example suggests that human-authored ones often choose less imageable words, (e.g., *"N"*), to meet strict phonetic similarity criteria, whereas LLMs, which are not specifically trained on phonetic data, tend to select more common and imageable words. Therefore, using LLM-generated keywords may reduce human effort on finding keywords for verbal cues that are not only phonetically similar but also imageable.

4.3.2 Verbal Cue

Table 2 compares verbal cues generated by our overgenerate-and-rank method against human-authored cues in Barron's book. We see that our method outperforms human authors on both imageability and coherence, especially on the latter, where human-authored verbal cues in Barron's book have much higher perplexity than LLM-generated ones. This result suggests that our method produces verbal cues that are not only vivid but also more coherent compared to Barron's book.

One possible reason for these results is that human-authored cues often make up unrealistic scenarios, thus significantly decreasing their coherence. For example, the highest perplexity score we observe is for the human-authored cue *"A polemical polar Mick call"* for the word *"polemical,"* scoring 3303.9. This cue evokes the image of someone in the Arctic, angrily using a phone. The phrase *"polemical polar Mick"* is a highly unusual combination; in contrast, the LLM-generated cue, *"The polemical John, standing like a pole, accuses me of having ideas as dark as coal,"* creates a more natural scenario with a more coherent sentence, resulting in a much lower perplexity score of 88.6. We provide a more detailed, per-word analysis in Section 6 on LLM-generated cues and also discuss feedback from real English learners.

Method		IMR$\uparrow$	PPL$\downarrow$
Keyword	Verbal Cue		
Barron	Llama3$_{FT}$	0.53	482.0
	GPT-4$_{Ours}$	0.54	147.1
Ours	Llama3$_{Ours}$	0.58	257.7
	GPT-4$_{Ours}$	0.61	156.4

Table 3: Ablation study changing our two-stage pipeline and the underlying LLM, evaluated on ImageReward and Perplexity. GPT-4$_{Ours}$ denotes our method.

4.4 Ablation

We perform an ablation study to assess the impact of several aspects of our verbal cue generation method. First, we assess the effectiveness of a two-stage pipeline, i.e., not generating keywords and using human-authored ones for verbal cue generation instead. Second, we evaluate the performance of a smaller, open-source language model, Llama 3-8B (Touvron et al., 2023). For this model, we conduct two experiments: fine-tuning with Barron's book (**FT**), and prompting with the same prompt as our method (**Ours**). See Supplementary Material D for the prompt and detailed model configuration.

Table 3 shows the ablation study results. We see that using our method to generate sets of keywords leads to much higher imageability in downstream verbal cue generation, compared to using human-authored keywords from Barron's book. We note that human-authored keywords do result in slightly better coherence, indicating a trade-off, however the significantly better imageability tilts the balance towards LLM-generated keywords. Meanwhile, we also see that Llama 3 results in lower performance on both metrics compared to GPT-4, even after fine-tuning on human-authored verbal cues. Notably, Llama3$_{Ours}$ successfully followed the prompt instructions only 58% of the time (either missing keywords or the target word), whereas Llama3$_{FT}$ followed them 90% of the time. This result suggests that automated KM generation is perhaps a task that is too difficult for smaller LMs to do well on, especially without fine-tuning on real KMs.

5 Human Evaluation

5.1 Setup

To further evaluate the imageability and coherence of KMs, and more importantly, their usefulness to language learners, we conduct human evaluation from both teachers' and learners' perspectives. The evaluation uses the same randomly selected target
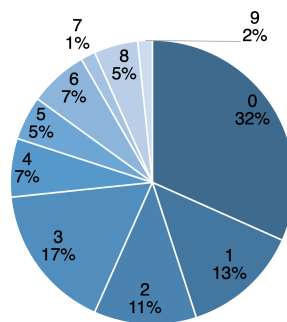

Figure 2: Unfamiliarity levels of the 60 target words among nine students in our evaluation. Each slice indicates the percentage of words where certain number of students indicate that they are unfamiliar with them. For example, "0, 32%" means that for 32% of the words, none of the students indicated unfamiliarity.

words described in Section 4.1. From the teachers' perspective, we employ four evaluators recruited through Upwork Inc. (2023), all of whom have experience teaching English at the high school level or preparing learners for English exams. From the learners' perspective, we hire nine university students through on-campus recruiting, including four freshmen and five sophomores; we require them to have recently took the SAT exams since our target words are SAT-level.

We conduct a preliminary survey among nine students that ask them to indicate their unfamiliarity with the 60 target words in our evaluation. The word *"polemical"* is identified as the most challenging, with all nine students unfamiliar with it. Similarly challenging words include *"abstemious,"* *"quiescence,"* and *"threadbare."* On the other end of the spectrum, 32% of all target words, such as *"aesthetic"* and *"authoritarian"*, are familiar to all students. This result shows the inherent difficulty of our task since, despite having five native English speakers out of nine in our group, some words are challenging even to these college-level individuals, justifying our participant selection criterion.

The experiment was conducted online through a web application. Prior to the experiment, we provided teachers and learners with a scoring rubric in Section C to calibrate their judgment, given the likely subjectivity among human evaluators on our evaluation criteria, especially usefulness. We ask evaluators to simultaneously rate the three criteria.

During the evaluation, the web application showed both LLM-generated and human-authored cues, one at a time. The ordering of target words and the source of the verbal cues (LLM vs. human)

were randomized and not disclosed to minimize potential biases.

5.2 Results

Table 4 shows the average of 5-point Likert scale ratings and the Spearman's rank correlation coefficient (Spearman's ρ), among teachers and learners, for both LLM-generated and human-authored verbal cues. We see that overall, participants found verbal cues, especially LLM-generated ones, to be imageable and coherent but relatively less useful. A Wilcoxon signed-rank test indicates a statistically significant difference (labeled as * with $p < 0.05$ in Table 4), which shows that LLM-generated cues are preferred over human-authored ones in all cases. We use the Wilcoxon signed-rank test instead of the t-test, as the ordinal nature of Likert scale data does not allow for the assumption of normal distribution. These findings align with those from the automated evaluation (Table 2), where our overgenerate-and-rank method also score better on both imageability and coherence than human-authored ones.

From the table, we make two observations: First, teachers show higher inter-rater agreement than learners, as indicated by pearman's ρ. This result can be explained by several key differences between the two participant groups. Teachers, with their professional background in education, tend to have a deep understanding of language, real-world pedagogical settings in language learning, and characteristics of many learners they have interacted with. Therefore, their assessments are likely less subjective since they focus on the pedagogical value of KMs across an entire learner population and the broader application of these cues within curricular goals. In contrast, learners, with less language proficiency, are primarily influenced by their personal background and experiences, which can lead to high subjectivity when individual preferences differ. We also find a lower level of agreement, with a much wider range of scores, on LLM-generated cues than on human-authored ones. This result can be explained by the diverse styles among LLM-generated cues spurring a high degree of subjectivity, since LLMs are trained on web-scale textual data, while the book contains verbal cues authored a small group of authors, with similar style.

6 Case Study

We now qualitatively analyze cues generated by our method based on feedback collected from learners through a post-experiment survey. We asked learners to identify and explain their top five most useful and least useful cues. From their responses, we discuss four representative examples of the most and least useful verbal cues, as shown in Table 5: the more useful cue being more imageable and coherent (higher IMR, lower PPL), more imageable but less coherent (higher IMR, higher PPL), more coherent but less imageable (lower PPL, lower PPL), and less imageable and coherent (lower IMR, higher PPL). We also show an image $\hat{i}$ used to calculate the IMR score in each case.

Highly Imageable & Coherent The keywords for *"artisan"* blend seamlessly into the verbal cue, creating an imageable scene, whereas the keywords for *"peripheral"* struggle to blend into the cue due to their complexity, which arises from the challenge of finding phonetically similar words that also match the syllable count. If this condition is relaxed, the verbal cue can improve: for *"pear, for, all,"* the cue *"A pear tree on the town's peripheral bears fruits for all."* blends seamlessly in the cue, achieving an IMR score of 1.39 and a PPL score of 184.8. This example shows the importance of selecting keywords that naturally fit the cue, rather than merely focusing on phonetic similarity and syllable count. Learners generally find cues that are more imageable and coherent to be more useful.

Imageable but not Coherent For *"exhaustive,"* the cue is useful because the keywords *"horse"* and *"stiff"* depict a vivid scene where a horse becomes tired after training. However, the phrase *"horse turned stiff"* is not commonly used (usually referred as *"stiff horse"*), leading to a higher PPL score. For *"phenomena,"* even though the verbal cue is coherent and consists of simple keywords, the keywords do not contribute to depicting a concrete scene nor relate to the word's meaning. This example suggests that the keywords should be closely linked to the word's meaning, even if it slightly impacts the cue's coherence.

Coherent but not Imageable For *"retract,"* the cue is useful because it creatively describes a detective retracting his suspicion after reviewing a clean record, but the cue for *"intimidate"* lacks creativity. Instead, the human-authored cue *"An intimate date tends to intimidate her."* is rated higher by learners across three aspects (imageability: 3.7 vs. 3.0, coherence: 3.8 vs. 3.0, usefulness: 3.9 vs. 2.4), evoking strong emotions to make it memorable. It is worth noting that the text-to-image model somewhat failed to describe the scenario in the cue for

	5-point Likert Scale				Spearman's ρ			
	Teachers		Learners		Teachers		Learners	
	Ours	Barron	Ours	Barron	Ours	Barron	Ours	Barron
Imageability	3.54* (1.25)	2.77 (1.42)	3.50* (1.23)	2.52 (1.25)	0.24	0.50	0.15	0.32
Coherence	3.60* (1.25)	2.90 (1.47)	3.33* (1.25)	2.58 (1.24)	0.34	0.56	0.27	0.35
Usefulness	2.89* (1.27)	2.20 (1.36)	3.11* (1.34)	2.29 (1.30)	0.31	0.59	0.20	0.32

Table 4: Comparison of mean (standard deviation) of 5-point Likert scale ratings and average Spearman rank-order correlation on verbal cues across different groups (four teachers and nine learners).

Most Useful				Least Useful			
Verbal Cue $\hat{\mathcal{V}}$	Image $\hat{i}$	IMR$^\uparrow$	PPL$^\downarrow$	Verbal Cue $\hat{\mathcal{V}}$	Image $\hat{i}$	IMR$^\uparrow$	PPL$^\downarrow$
In the *art*-loving town, the **artisan** sips his *tea* under the *sun*.		1.74	86.0	A *pair* teased with 'what *if her doll* disappears?', hiding it in **peripheral** areas, but her attention captured it all.		0.08	380.7
An *ex*-champion *horse* turned *stiff* from his **exhaustive** training.		1.13	358.7	Sarah, studying **phenomena**, was on the *phone* when she had to say, "*No, Ma*."		0.07	95.3
After *reading* the clean *track* record of the suspect, the detective had to **retract** his suspicion.		-0.11	71.1	*In* a *tea* gathering, a *mate* walked in to **intimidate** everyone.		0.51	515.3
Worn to a *thread*, the *bear* became **threadbare** but still cherished.		0.44	91.2	*Veer*'s talent was *too* remarkable, *so* he became a **virtuoso**.		1.70	80.0

Table 5: Most and least useful verbal cues (keywords in *italic* and a target word in **bold**) indicated by learners. Colors indicate whether a score is higher (teal) or lower (magenta).

"retract," leading to a lower IMR score. This example suggests that certain properties of verbal cues, such as creativity or emotion, may overcome other problems such as a lack of imageability.

Less Imageable & Coherent For *"threadbare,"* the cue is useful because the context *"worn to a thread"* successfully conveys the meaning of the target word. The cue for *"virtuoso"* lacks an appropriate context, failing to indicate any connection to art. Regardless of the relevance of other generated cues containing art contexts like pianist or violin, they are not highly ranked due to their excessive length, illustrating a trade-off between suggestions ii) and iv) from the pilot study. Although learners think the verbal cue for *"threadbare"* is both imageable and coherent, the metrics indicate otherwise, likely due to the abstract nature of *"cherish,"* which is challenging to visualize and the ambiguity surrounding whether *"bear"* refers to a teddy bear.

The discrepancies highlighted above between usefulness ratings and imageability/coherence of the verbal cue suggest that future work should focus on aligning automated metrics with human preferences, possibly using preference optimization (Rafailov et al., 2024), and improv-

ing the fidelity of verbal-visual conversion using techniques developed for abstract linguistic metaphors (Chakrabarty et al., 2022).

7 Related Work

Imageability is defined as *"the ease with which a word arouses sensory images"* (Paivio et al., 1968). To quantify this intangible aspect of language, psycholinguists and psychologists have compiled human imageability ratings databases like the MRC Psycholinguistic Database (Wilson, 1988) and Glasgow Norms (Scott et al., 2019). However, collecting this data through interviews is costly and time-consuming, making it difficult to scale psycholinguistic databases to the size of modern NLP corpora like the Corpus of Contemporary American English, which includes over 60,000 lemmas with frequency and parts of speech data (Wu and Smith, 2023).

Recent studies have addressed the challenge of scalability in assessing imageability by applying machine learning techniques to automate the collection of imageability ratings. These efforts include predicting imageability using supervised learning (Ljubešić et al., 2018a), and employing im-

age data mining to estimate word imageability by analyzing a various visual features (Kastner et al., 2020). Moreover, advancements have extended these predictive models to the sentence level. For instance, researchers have developed methods to evaluate the visual descriptiveness of captions and introduced ways to calculate a sentence's imageability score based on the imageability scores of its constituent words (Umemura et al., 2021). Additionally, computational techniques have been proposed that utilize text-to-image models to generate images and measure sentence imageability by calculating the cosine similarity between the image and word embeddings (Wu and Smith, 2023).

8   Conclusions and Future Work

In this paper, we explored using large language models for the task of automated keyword mnemonics generation for vocabulary learning, via a novel overgenerate-and-rank method. Through both automated and human evaluation with both English teachers and learners, we found that our generated verbal cues are comparable to or better than human-authored ones, in terms of imageability, coherence, and usefulness. We also studied the intrinsic subjectivity among learners in our evaluation through qualitative feedback, which led to relatively lower inter-rater agreement.

There are many avenues for future work. First, due to this intrinsic subjectivity, we need to develop methods to generate *personalized* cues that each individual learner will find useful, by adapting to their personal language knowledge and cultural background. Second, we need to study automated *visual* cue generation, in the form of images or even videos, and evaluate their quality separately from the verbal cues. Third, we plan to conduct an experiment with real language learners in classrooms, to evaluate whether our automatically generated verbal and even visual cues can indeed enhance vocabulary recall over a long period of time. Fourth, for each language, there may be better ways of generating richer mnemonics by leveraging linguistic and cultural nuances. For instance, in Mandarin, the character "休" ("to rest") can be linked to the keyword "shoe" based on sound "xiū", but a more meaningful mnemonic could involve the character's components—"人(person)" and "木(tree)"—that combine to form its meaning. Developing such language-specific verbal cues is a promising direction for future work.

9   Acknowledgement

Limitations

The study has a few limitations that should be considered. First, the scope of the study on Keyword Mnemonics is limited to English. Future studies could benefit from applying to second language acquisition. Second, a limited number of teachers and learners evaluating the verbal cues generated by the LLM might not be enough to gain a comprehensive understanding of their quality and usefulness. The opinions and insights of a larger pool of English experts would provide a broader range of perspectives and expertise, contributing to a more robust evaluation of the LLM-generated cues. Third, the study primarily assessed KM usefulness through teacher and learner preference ratings rather than practical application on long-term memory tests. While teacher and learner opinions provide valuable insights, a more robust evaluation would involve measuring the impact of KMs on actual long-term memory retention in language learners. Incorporating such practical assessments would accurately reflect the KMs' efficacy in real-world language learning scenarios.

Ethics Statement

Our human experiment was conducted with the approval of the Institutional Review Board (IRB). Prior to their participation, we provided teachers and learners with a consent form that thoroughly outlined the potential risks, benefits, time commitment, expected actions, and compensation. The compensation offered to teachers and learners adhered to the recommended amount for academic research studies, which is no less than the federal minimum wage of $7.25 per hour.

References

International Phonetic Association. 1999. *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet.* Cambridge University Press.

Richard C Atkinson and Michael R Raugh. 1975. An application of the mnemonic keyword method to the acquisition of a russian vocabulary. *Journal of experimental psychology: Human learning and memory*, 1(2):126.

Nishant Balepur, Matthew Shu, Alexander Hoyle, Alison Robey, Shi Feng, Seraphina Goldfarb-Tarrant, and Jordan Boyd-Graber. 2024. A smart mnemonic sounds like" glue tonic": Mixing llms with student feedback to make mnemonic learning stick. *arXiv preprint arXiv:2406.15352*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.

Sam Burchers, Max Burchers, and Bryan Burchers. 2000. Vocabulary cartoons ii: Sat word power.

Tuhin Chakrabarty, Arkady Saakyan, Olivia Winn, Artemis Panagopoulou, Yue Yang, Marianna Apidianaki, and Smaranda Muresan. 2022. I spy a metaphor: Large language models and diffusion models co-create visual metaphors. *OpenReview Preprint*. Preprint under review.

Boris Dayma, Suraj Patil, Pedro Cuenca, Khalid Saifullah, Tanishq Abraham, Phúc Le Khac, Luke Melas, and Ritobrata Ghosh. 2021. Dall· e mini. *HuggingFace. com. https://huggingface. co/spaces/dallemini/dalle-mini (accessed Sep. 29, 2022)*.

Philip Geer and Susan Geer. 2018. *Picture These SAT Words!* Barron's Educational Series, Hauppauge.

James W Heisig. 2011. *Remembering the kanji 1: A complete course on how not to forget the meaning and writing of Japanese characters*. University of Hawaii Press.

Marc A Kastner, Ichiro Ide, Frank Nack, Yasutomo Kawanishi, Takatsugu Hirayama, Daisuke Deguchi, and Hiroshi Murase. 2020. Estimating the imageability of words by mining visual characteristics from crawled image data. *Multimedia Tools and Applications*, 79(25):18167–18199.

Anisia Katinskaia and Roman Yangarber. 2021. Assessing grammatical correctness in language learning. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 135–146.

Koohii Kanji. 2024. Koohii kanji. https://kanji.koohii.com/.

Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. Age-of-acquisition ratings for 30,000 english words. *Behavior research methods*, 44:978–990.

Jaewook Lee and Andrew Lan. 2023. Smartphone: Exploring keyword mnemonic with auto-generated verbal and visual cues. In *International Conference on Artificial Intelligence in Education*, pages 16–27. Springer.

Kai-Hui Liang, Sam Davidson, Xun Yuan, Shehan Panditharatne, Chun-Yen Chen, Ryan Shea, Derek Pham, Yinghua Tan, Erik Voss, and Luke Fryer. 2023. Chatback: Investigating methods of providing grammatical error feedback in a gui-based language learning chatbot. In *Proceedings of the 18th workshop on innovative use of NLP for building educational applications (BEA 2023)*, pages 83–99.

Nikola Ljubešić, Darja Fišer, and Anita Peti-Stantić. 2018a. Predicting concreteness and imageability of words within and across languages via word embeddings. In *Proceedings of the 3nd Workshop on Representation Learning for NLP, RepL4NLP@ACL 2018, Melbourne, Australia, July 20, 2018*.

Nikola Ljubešić, Darja Fišer, and Anita Peti-Stantić. 2018b. Predicting concreteness and imageability of words within and across languages via word embeddings. In *Proceedings of the 3nd Workshop on Representation Learning for NLP, RepL4NLP@ACL 2018, Melbourne, Australia, July 20, 2018*.

Mnemonic Dictionary. 2024. Mnemonic dictionary. https://mnemonicdictionary.com/.

Yuki Okano, Kotaro Funakoshi, Ryo Nagata, and Manabu Okumura. 2023. Generating dialog responses with specified grammatical items for second language learning. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 184–194.

Allan Paivio, John C Yuille, and Stephen A Madigan. 1968. Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of experimental psychology*, 76(1p2):1.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021. High-resolution image synthesis with latent diffusion models. *Preprint*, arXiv:2112.10752.

Manolis Savva, Angel X Chang, Christopher D Manning, and Pat Hanrahan. 2014. Transphoner: Automated mnemonic keyword generation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3725–3734.

Graham G Scott, Anne Keitel, Marc Becirspahic, Bo Yao, and Sara C Sereno. 2019. The glasgow norms: Ratings of 5,500 words on nine scales. *Behavior research methods*, 51:1258–1270.

5530

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*.

Gladys Tyen, Mark Brenchley, Andrew Caines, and Paula Buttery. 2022. Towards an open-domain chatbot for language practice. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 234–249.

Kazuki Umemura, Marc A Kastner, Ichiro Ide, Yasutomo Kawanishi, Takatsugu Hirayama, Keisuke Doman, Daisuke Deguchi, and Hiroshi Murase. 2021. Tell as you imagine: Sentence imageability-aware image captioning. In *MultiMedia Modeling: 27th International Conference, MMM 2021, Prague, Czech Republic, June 22–24, 2021, Proceedings, Part II 27*, pages 62–73. Springer.

Upwork Inc. 2023. Upwork: The world's work marketplace. Accessed: 2023-06-22.

Michael Wilson. 1988. Mrc psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior research methods, instruments, & computers*, 20(1):6–10.

Si Wu and David A Smith. 2023. Composition and deformance: Measuring imageability with a text-to-image model. *arXiv preprint arXiv:2306.03168*.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. 2024. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36.

Chak Yan Yeung and John Lee. 2021. Character set construction for chinese language learning. In *16th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2021)*, pages 59–63. Association for Computational Linguistics (ACL).

Mengyu Zhang, Weiqi Wang, Shuqiao Sun, and Weiwei Sun. 2021. Negation scope resolution for chinese as a second language. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–10.

Brian Zylich and Andrew Lan. 2021. Linguistic skill modeling for second language acquisition. In *LAK21: 11th International Learning Analytics and Knowledge Conference*, pages 141–150.

Supplementary Material

A   Prompt

Prompt	Game Name: Sounds Like A Word Game Objective: The objective of this linguistic game is to dissect a target word into syllables and find an array of common real words that, when pronounced, produce a similar sound to the target word. This game is not about spelling, but about phonetics. Game Rules: 1. Each part of the target word should correspond with a standalone word, not a prefix or a suffix. 2. The words selected should be suitable for those preparing for SAT, so players must choose words that are commonly used and understood at that level. This means that obscure, rare, or highly specific words should be avoided. 3. The goal is to find words that sound like the target word when pronounced together, not necessarily to match the exact number of syllables in the target word. Therefore, the number of keywords might be less or more than the number of syllables in the target word. For instance, "duplicity" with four syllables could translate into "do please city" with three keywords. 4. Players must correctly split each target word into syllables and locate common words that closely match these syllables in terms of sound. This game tests both phonetics knowledge and creativity. 5. Players should avoid using slang, offensive, or sexually suggestive words as keywords. The chosen words should be appropriate and respectful. Here are the examples of how the input and output should look like: [Input] Target word: alleviate [Output] Keywords: a, leaf, he, ate [Input] Target word: belie [Output] Keywords: bee, lie [Input] Target word: duplicity [Output] Keywords: do, please, city [Input] Target word: appease [Output]
Response	Keywords: a, peas

Table 6: Prompts for generating syllabic keywords.

Prompt	Game Description:
	In StoryWeave, players are given a target word and a set of keywords. The task for players is to craft an engaging story using these words cleverly. The ultimate challenge is to construct a narrative that not only incorporates the target word but also includes the keywords in the exact order presented. The beauty of the game lies in the players' use of imagination and language to unfold a thought-provoking plot and articulate characters.
	Upon completion of the narrative, players generate a summary of their narrative, with an emphasizing requirement: the keywords still must appear, but in the same sequence as provided in the original story. This twist lies in maintaining the faithfulness of the summary to the initial tale, allowing no room for distortion of the original plotline.
	In StoryWeave, creativity and logic flow in harmony, leading to captivating narratives that tie readers from a solitary word to an entire universe-shaped by strategic word placement and unrestricted imagination.
	Game Rules/Constraints: 1. Each player is handed a target word and a set of keywords. 2. The target word should be weaved into the story while the keywords are to appear exactly in the order given. 3. After crafting the narrative, a summary is to be produced. This summary will also utilize the keywords, keeping them in the identical order as featured in the initial story. 4. Players are strictly prohibited to rearrange the sequence of the keywords. 5. Both the full-length story and its summary should unequivocally reflect the original plotline, with no discrepancies allowed.
	The idea is to balance the ingenious use of vocabulary to construct a story and maintain absolute fidelity to the given keywords and their order. The essence of this game lies in the inventiveness of constructing an interesting narrative around the given 'word constraints'.
	Here are the examples of how the input and output should look like:
	[Input] Target word: alleviate Meaning: relieve; make more bearable Keywords: a, leaf, he, ate [Output] Story: He found himself famished, left with nothing but a single leaf on his plate. With a sigh of resignation, he ate it, hoping to alleviate his gnawing hunger a bit. Summary: On his plate, there was a leaf he ate to alleviate his hunger.
	[Input] Target word: belie Meaning: contradict; give a false impression Keywords: bee, lie [Output]

	Story: In the sunlit meadow, Sunbeam, a deceitful bee, told an enchanting lie. His sweet tales belied his venomous sting, misleading creatures who fell prey to his deceptive harmony. Summary: The bee tells a lie that belies its nasty sting. [Input] Target word: appease Meaning: soothe; relieve Keywords: a, peas [Output]
Response	Story: The young boy was upset, he hated vegetables. His mother, understanding his distaste, offered him a plate of peas. She hoped this would appease his frustration and coax him into eating healthier. Summary: To appease the boy's dislike for vegetables, his mother offered him a plate of peas.

Table 7: Prompts for generating verbal cues.

Prompt	What would be the most likely word in to be in the [MASK]? List 5 candidates. To [MASK] the boy's dislike for vegetables, his mother offered him a plate of peas.
Response	1. Overcome 2. Counteract 3. Combat 4. Challenge 5. Confront

Table 8: Prompts for generating candidates for context completeness ranking (R_{cont}).

B Verbal Cues

Human-authored cues are from Geer and Geer (2018), a work copyrighted by Barron's Educational Series, Inc. LLM-generated cues and human-authored cues are only used for human evaluation.

Word	LLM-generated	Human-authored
abstemious	*Ab*'s diet of only the *stem* of vegetables taught *me* and *us* to be more **abstemious**.	*Ab*'s *steamy* with *us* when he tells us to be **abstemious**.
accolade	The *echo* of the *lad*'s music earned him an **accolade**.	*A coal aide* receives an **accolade**.
aesthetic	*As the* clock *ticked*, he was captivated by the **aesthetic** appeal of the painting.	Oh beautiful *S*! *That tick* tickles the **aesthetic** sense!
appease	He cooks *a* dish with *peas* to **appease** her anger.	Tom was **appeasing** a pot *o' peas*.
archaic	The twins wonder "*are* these *cake* instructions?" from an **archaic** recipe book.	Sick kangaroos ride on **archaic** *Ark K* - "*Ick*!" is all they can say.
artisan	In the *art*-loving town, the **artisan** sips his *tea* under the *sun*.	"*Art is sin*," says the Puritan to the **artisan**.
ascendancy	The *ass* and *hen dance* by the *sea* to determine their **ascendancy**.	The Egyptians are doing the *Ascend Dance* to **ascendancy**.
authoritarian	The citizens of *Ought* face the choice to submit *or* resist the **authoritarian** regime, causing their freedom to *tear* like *rain*.	**Authoritarian** *Arthur Tarian* tells Ian to tear up other authors' works.
beguile	A *bee* and a *guy* **beguile** people in each supermarket *aisle* they walk down together.	Be *Guy* well, or *be Guy ill*, he must **beguile** the nurse from Brazil.
censure	John's *sin* made his mother *sure* to **censure** him.	A *cent* with a *sore* **censured** at the Century Bank.
compliance	Using a *comb* to *ply* his hair, Billy observes *ants'* **compliance** to their queen's orders.	"*Come*, *ply ants* with honey to gain their **compliance** with our plan."
concede	The *con* man was outsmarted by a humble *seed* seller, leading him to **concede** defeat.	"I **concede** that it's *cons' seed*, but I need feed!"
conciliatory	A *con* man helps a wounded *seal he* found, *tearing* his clothes for a bandage in a **conciliatory** act, leading to his *re*formation.	The *con's silly story* - a **conciliatory** gesture.
contend	The city's *con* artist *tends* to **contend** that he's misunderstood.	He **contends** that a *con tends* to return.
conventional	*Con* finds a *van*, and against his **conventional** nature to *shun all* rules, reports it.	At a *convention*, *all* must wear **conventional** suits.
conviction*	*Con* artist *Victor*'s **conviction** was proven by the necklace found against his *shin*.	"It is my **conviction** that every *convict* should *shun* guns."
cordial	In the city's *core*, the **cordial** shopkeeper makes fair *deals*.	"I've cut the *cord*, *Jill*!" (Relations are no longer **cordial**.)

cryptic	In the *crypt*, a mysterious '*tick*' sound was a **cryptic** secret.	A **cryptic** *crypt tick*.
degradation	The *degrade* endured by the *Asian* immigrant led to his personal **degradation**.	**Degradation** from a "*D*" *grade date - shun* it!
depravity	The *deep*, *rave*-filled *city* hides a world of **depravity**.	*Deep* in the *rabbit* warren, *he* tells them a tale of **depravity**.
deprecate	Luna looks at the *deep* sea *wreck* with *hate*, **deprecating** man's recklessness.	Near *Deep-Wreck 8*, they **deprecate** the "Catch of the Day."
disputatious	At "*Dis* Church", the *pew* is filled with **disputatious** locals, until Mrs. *Tat* attempts to *shush* them.	*Dispute 8. Just* another dispute between **disputatious** dates.
divergent*	The *ant*, on the *verge* of *dying*, faced **divergent** paths.	*Di* on the *verge* of going with a *gent* on a **divergent** path.
egotism	Mr. *Ego*, over his *tea*, tallies up the *sum* of his achievements, revealing his **egotism**.	The main concern of *E goat is himself.* What **egotism**!
emulate	The usually punctual *emu* was *late*, yet other birds still sought to **emulate** its habits.	"*Em*, *you*'re *late*! Must you **emulate** girls who make their dates wait?"
enmity	The *hen* displays her **enmity** towards the farmer's golf *mitt* at the *tee*.	The *N mitt he* wears causes **enmity**.
exhaustive	An *ex*-champion *horse* turned *stiff* from his **exhaustive** training.	"Our **exhaustive** battery of tests does not *exhaust Steve*."
feasible	With the rise in *fees*, John was still *able* to pay, making it a **feasible** solution for him.	It's not **feasible** to pay *fees* to *a bull*.
flagrant	The king's **flagrant** *flaw* was to *grant* favors indiscriminately, causing outrage.	**Flagrant** hostility at a *flag rant*.
gullible	The *gull*, by moving his *lip*, convinces the **gullible** *bull* of his stories.	"If you believe in the *gull-a-bull*, you must be **gullible**!"
ignominy	The man with an *egg* has *no money*, embodying his **ignominy**.	*A gnome, Minnie*, suffers no **ignominy**.
illusory	*Ill* and fearing he will *lose* everything, John says *sorry* in his **illusory** world.	*A Lew*, *sorry* for having followed an **illusory** dream.
implement	An *imp*, after a deep *lament*, decides to **implement** changes to his behavior.	They **implement** a curfew to an *imp lament*.
inclusive	*In* the search for a *clue*, the detective's **inclusive** method acted like a *sieve*, filtering through all the evidence.	Look *in clues if* you seek **inclusive** evidence.
inconsequential	The artist finds *ink on* his *sequence* of strokes but remains *chill*, deeming it **inconsequential**.	"It's **inconsequential** how we go, but *in con sequence shall* we go if you insist!"
incorrigible	The artist used *ink* for his drawing of an *or*-like *ridge* and a *bull*, but the bull was an **incorrigible** mistake.	Trying to *encourage a bull* to cease his **incorrigible** ways.
indifferent	He was **indifferent**, even *in a different* situation.	People *in different* lands being **indifferent** to each other.

ingenious	Dr. Jensen found an **ingenious** solution *in* the *gene*, saying '*yes*' to his eureka moment.	*In genius* we find **ingenious** ideas.
intimidate	"*In* a *tea* gathering, a *mate* walked in to **intimidate** everyone."	An *intimate date* tends to **intimidate** her.
intrepid	*In* his *trip*, the **intrepid** explorer fearlessly faced what *hid* in the shadows.	Even the most **intrepid** explorer should, *in* his *trip*, *heed* warnings.
oblivion	Joe wished, "*Oh, live beyond* this hardship," but his dreams sank into **oblivion**.	*Ob lived* and *eon* before Oblantis fell into **oblivion**.
opportunist	Always looking *up* for a *port* of advantage, *you* would see an **opportunist** building his *nest* on others' misfortunes.	At *a port* near *Tunis*, an **opportunist** waits.
opulence	An overwhelmed man exclaims "*Oh*!" as he sits at a *pew*, observing the cathedral's **opulence** through his *lens*.	"An *opal lance*? Such **opulence** in this palace!"
peripheral	A *pair* teased with 'what *if her doll* disappears?', hiding it in **peripheral** areas, but her attention captured it all.	"There's a *pear for all* on the **peripheral** pear trees!"
phenomena	Sarah, studying **phenomena**, was on the *phone* when she had to say, "*No, Ma*".	"And now its a *fin omen - ahhh*! - rare **phenomena**"
polemical	The **polemical** John, standing like a *pole*, accuses *me* of having ideas as dark as *coal*.	A **polemical** *polar Mick call*.
quiescence	The *key* to peace is in the *essence* of **quiescence**.	*Qwee Essence* brings **quiescence**.
rant	She *ran* out of patience waiting for her *tea* and went on a **rant**.	"By *Ra*, that *ant* can **rant**!"
ratify	The *rat* agrees to **ratify** a *tie*-up treaty for a *fee*.	"I'll be a *rat if I* **ratify** this treaty."
recount	Mathematician *Ree* had to *count* and then **recount** to secure her win.	Recounting how *Rick* was picked to *count* the votes in the **recount**.
rectify	Wearing his *red tie*, he decided to *fie* and **rectify** his mistake.	"I'll be *wrecked if I* don't **rectify** my neighbor's behavior."
rescind	The courier had the letter on his *wrist* to *send*, but the order was **rescinded**.	*Reese sent* his team a memo to **rescind** his previous one.
retract	After *reading* the clean *track* record of the suspect, the detective had to **retract** his suspicion.	*Rhet tracked* through the woods to **retract** his words.
rigor	On the *ridge*, John says "*Er...*" but the **rigor** of his training encourages him to continue.	The *rig* is checked by captain *Gore* with **rigor**.
stoic	A fire starts from the *stove*, to which the **stoic** chef merely reacts with '*ick*'.	**Stoic** customers tuck at the *Stowe Wick* in Stowe, Vermont.
surreptitious	'*Sir*', a '*rep*' manager, needs to hide a scandalous *titbit* with a **surreptitious** '*shush*'.	*Sir Repetitious* in a **surreptitious** operation.
tantamount	The *ant* finds a *tea mound* **tantamount** to a valuable treasure.	Using *Tant* as *a mount* to ascend El Pico in Peru is **tantamount** to saying alpacas' rights are few!

threadbare	Worn to a *thread*, the *bear* became **threadbare** but still cherished.	*Thread Bear* takes orders from the **threadbare** customers.
unwarranted	An unexpected '*un*' *war ran* through *Ted*'s land, leading to his **unwarranted** accusation.	"*A warrant*?" *Ted* asked. "That's **unwarranted**."
virtuoso	*Veer*'s talent was *too* remarkable, *so* he became a **virtuoso**.	This **virtuoso** has a *virtue oh so* rare - he spreads cheer far and near.

Table 9: Verbal cues used for human evaluation. Keywords are represented in *italic*, while a target word is in **bold**. * indicates an anomaly in verbal cue generation using LLM.

In the case of the target word "conviction," only one set of keywords ("con," "vict," "shun") were generated. In the case of the target word "divergent," the model failed to arrange the keywords correctly, which should have shown in the order of ("die," "verge," "ant").

C Human Evaluation

C.1 Web Interface

Instructions

Imagine yourself as a student who is learning the word for the first time.

Evaluate the sentence based on the following criteria.

Current progress: 5 / 60
Time spent: 36s
Average time: 21s

Word (Keywords)	Illusory (ill, lose, sorry)
Definition	deceptive; not real
Sentence	*Ill* and fearing he will *lose* everything, John says *sorry* in his **illusory** world.

Coherence
RATING GUIDE ⌄
Medium
★ ★ ★ ☆ ☆

Imageability
RATING GUIDE ⌄
Medium
★ ★ ★ ☆ ☆

Useful
RATING GUIDE ⌄
Medium-High
★ ★ ★ ★ ☆

SUBMIT

Figure 3: Web application interface for human evaluation.

C.2 Criteria

Guidelines for rating 5-point Likert Scale on imageability, coherence, and usefulness.

Scale	Explanation
High (5)	The sentence evokes a vivid and detailed mental image, making it easy to visualize the scene or situation described in the sentence.
Medium (3)	The sentence evokes a reasonable level of imagery: minor inconsistencies may exist in the description, but a mental image can still be formed.
Low (1)	The sentence lacks substantial imagery, making it challenging to form any meaningful mental image.

Table 10: Instructions for rating **imageability** of the verbal cues.

Scale	Explanation
High (5)	The sentence is highly coherent: the meaning is clear, and the wording is natural.
Medium (3)	The sentence is moderately coherent: minor issues affect clarity in understanding the meaning.
Low (1)	The sentence lacks coherence: it's difficult to understand or read because it is illogical or grammatically incorrect.

Table 11: Instructions for rating **coherence** of the verbal cues.

Scale	Explanation
High (5)	The sentence given is a useful tool for memorizing the vocabulary word. It will help me remember the meaning of the word, or I imagine it would be helpful to others.
Medium (3)	The sentence provided has issues that affect how useful I find it, but with some minor modifications, it could be useful.
Low (1)	This sentence is not useful at all.

Table 12: Instructions for rating **usefulness** of the verbal cues.

D   Ablation Study

D.1   Llama3 Fine-tuning

The book "Picture These SAT Words!" includes a total of 314 words. For our study, we allocated 60 of these words to a test set for human evaluation. On the remaining words, we perform an 80-20 train-validation set split to fine-tune Llama3 for the KM generation task. We excluded the words "alleviate" and "belie" from this training set as they were used as in-context examples. We leverage an early-stopping approach where we evaluate the model on the validation set after each epoch and utilize the model weights with the highest validation performance at test time.

Specifically, we utilize the `meta-llama/Meta-Llama-3-8B` as the base model, paired with the Adam optimizer, set to its default parameters (b1=0.9, b2=0.999, eps=1e-6). We utilize an learning rate of 2e-5 and batch size of 2 for 5 epochs. Instead of full fine-tuning we use, LoRA adaptors for the modules `q_proj`, `k_proj`, `v_proj`, and `o_proj`, with each adaptation characterized by a rank r of 16, alpha of 16, and dropout rate of 0.1.

D.1.1   Prompt

Prompt	Your task is to create memorable keywords and a sentence that helps memorize a specific target word. The keywords should consist of words that resemble the phonetic sounds of the target word's syllables. The sentence should incorporate the provided keywords and the target word. Please adhere to the following rules: 1. Keep the original keyword order in the sentence. 2. Provide clear context for the target word. 3. Use words at the same (complexity) level as or lower than the target word. 4. Keep the cue short; long ones are not helpful. [Input] Target word: alleviate Meaning: relieve; make more bearable Keywords: a, leaf, he, ate [Output] Sentence: On his plate, there was a leaf he ate to alleviate his hunger. [Input] Target word: belie Meaning: contradict; give a false impression Keywords: bee, lie [Output] Sentence: The bee tells a lie that belies its nasty sting. [Input] Target word: appease Meaning: soothe; relieve Keywords: a, peas [Output]
Response	Sentence: To calm the toddler, a peas dish was made to appease her cries.

Table 13: Prompt for ablation Llama3$_{\text{FT}}$.

D.1.2   Llama3 Sampling Configuration

For both fine-tuning and zero-shot, we perform nucleus sampling of Llama3. We set a `temperature` of 1.0, ensuring a balanced approach to novelty and feasibility in outputs. We set `top_p` to 0.95, which

allows the model to consider a range of token possibilities, enhancing creativity without straying too far from plausible completions. Furthermore, `top_k` is restricted to 50, focusing the model's choices to the top 50 most probable next tokens, which helps in maintaining coherence and relevance in the generated text.

E Software Package

In our study, we employed various software packages. We used `spearmanr` and `wilcoxon` from `scipy.stats` for calculating Spearman's rank correlation coefficient and performing the Wilcoxon signed-rank test, respectively. For word lemmatization, the `nltk` package was utilized. Additionally, we used `lmppl`, specifically `meta-llama/Meta-Llama-3-8B`, for computing perplexity.