SRS-Stories: Vocabulary-constrained multilingual story generation for language learning

Wiktor Kamzela[1] **and Mateusz Lango**[1,2] **and Ondřej Dušek**[2]
[1]Poznan University of Technology, Institute of Computer Science, Poznan, Poland
[2]Charles University, Faculty of Mathematics and Physics, Prague, Czechia
wiktor.kamzela@student.put.edu.pl, {lango,odusek}@ufal.mff.cuni.cz

Abstract

In this paper, we use large language models to generate personalized stories for language learners, using only the vocabulary they know. The generated texts are specifically written to teach the user new vocabulary by simply reading stories where it appears in context, while at the same time seamlessly reviewing recently learned vocabulary. The generated stories are enjoyable to read and the vocabulary reviewing/learning is optimized by a Spaced Repetition System. The experiments are conducted in three languages: English, Chinese and Polish, evaluating three story generation methods and three strategies for enforcing lexical constraints. The results show that the generated stories are more grammatical, coherent, and provide better examples of word usage than texts generated by the standard constrained beam search approach.

1 Introduction

One of the very popular ways to learn a foreign language is to use a Spaced Repetition System (SRS) (Leitner, 1972; Reddy et al., 2016) such as Anki (Elmes, 2015), SuperMemo (Woźniak and Gorzelańczyk, 1994) or HackChinese.[1] SRS is essentially an application for learning new vocabulary by reviewing a deck of flashcards. Every day, SRS selects a set of flashcards containing words learned in the past but predicted by the algorithm to be in danger of getting forgotten by the user in the near future. In this way, the system optimises the user's ability to recall all the vocabulary they have learned, while minimising the number of repetitions. Although this strategy has proved effective in teaching the user to recall new words, it is of limited value for real vocabulary acquisition (Aslan, 2011; Miles and Ehri, 2017). Words are reviewed out of context and learners often find it difficult to use them naturally in a sentence. In addition, systematically reviewing flashcards with new vocabulary can become tedious.

Graded readers are resources that teach new vocabulary in context and are more interesting for the learner (Nation and Wang, 1999). These are books written with simplified vocabulary and with more difficult words accompanied by translations or definitions. The more difficult words are usually used several times in the story to help the reader remember them. While the advantages of this approach are clear, it may be difficult to find a graded reader at the appropriate language level (especially for languages other than English), and the learner has little control over the vocabulary studied. The repetition of learned words is not optimized in any way and learned vocabulary may be forgotten (Waring and Takaki, 2003).

In this paper, we explore the possibility of automatically generating graded readers (stories) that use vocabulary limited to words known to the user. In addition, the content of the story is specially designed to allow the introduction of selected new words, which are used several times in the story with meaningful context. The text can also deliberately include words that the user needs to review. Since such an approach can be coupled with a SRS system, effectively replacing flashcards with enjoyable stories, we call it SRS-Stories.

Our contributions are as follows:

- We propose three prompting strategies for large language models to generate SRS-Stories, i.e. coherent texts written to teach words arbitrarily selected by SRS.
- We design three strategies for enforcing vocabulary constraints based on text rewriting, involving an external non-neural constraint verifier and iterative story re-generation.
- We carry out an experimental evaluation of the proposed story generation techniques at different language levels in three languages: English,

[1]http://www.hackchinese.com

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 630–645
November 4-9, 2025

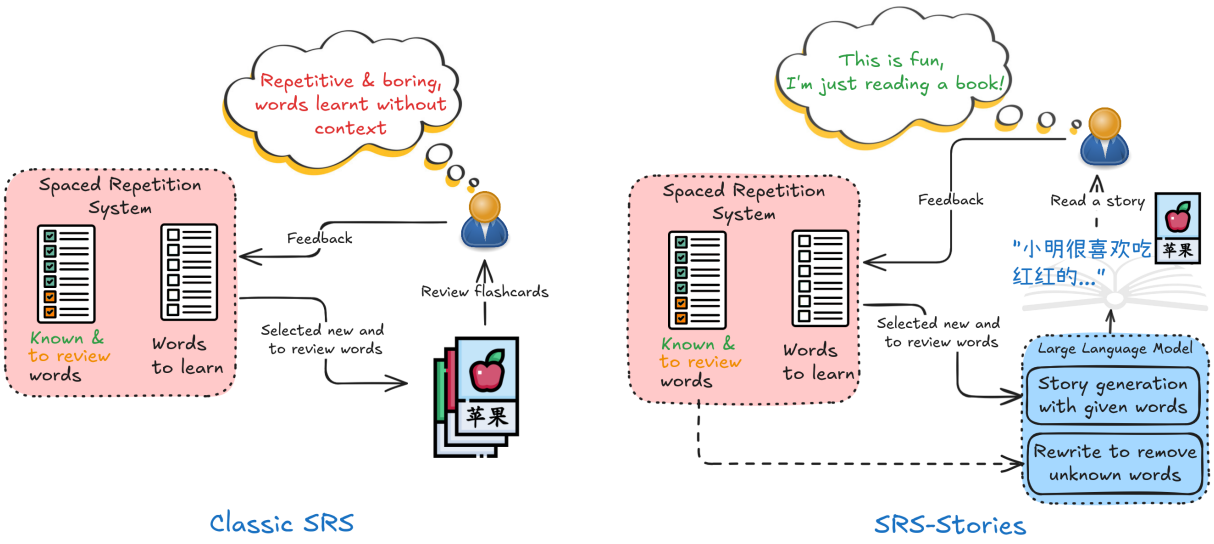

Figure 1: Overview of a classic SRS system and the presented approach: SRS-Stories.

Chinese and Polish.

The results show that the proposed prompting strategies generate stories that are more interesting, coherent and grammatically correct than those generated by constrained beam search (CBS; Post and Vilar, 2018; Hu et al., 2019), a classic lexical constraint enforcement method.

2 SRS-Stories

SRS-Stories is a new language learning system that combines the benefits of graded readers with intelligent SRS vocabulary review schedules (see Fig. 1 for an overview of the proposed approach). The system consists of a standard SRS engine and a large language model (LLM) with story generation capabilities. The SRS engine stores the list of new words that the user wants to learn, as well as the list of previously learned words. For each learning session, the SRS estimates which words need to be reviewed and picks a set of new words to learn. Based on the selected words, the LLM generates a coherent story written with vocabulary appropriate to the learner.

The story contains the words selected by the SRS, allowing the user to review them seamlessly as they read. In particular, new words are used multiple times in meaningful contexts in the generated text, naturally providing the user with examples of how the word can be used in sentences.

2.1 Task

The task considered in this paper is to generate a story under certain lexical constraints. Given a set of words V known by the user and the words to learn L selected by SRS, generate a coherent story $X = \{x_1, x_2, ..., x_n\}$ such that all words are suitable for the user ($x_i \in V \cup L$) and the new words appear at least c times ($\forall_{l \in L} \sum_{x \in X} [\![x = l]\!] \geq c$). In this paper we assume $c = 3$.

2.2 Methods

We propose three prompting strategies for generating a story that uses the selected words the required number of times. These can be combined with three approaches for enforcing lexical constraints, which rewrite generated stories using only appropriate words ($V \cup L$, i.e., previously known or selected new ones). See Appx. B for full prompts.

2.2.1 Story generation strategies

Simple Prompting We prompt the model to write a story of a certain length (in our experiments, 500-750 words) that uses selected words at least three times in sentences that clearly show their typical usage.

Planning This method aims to generate more interesting and coherent stories with given words by dividing the story generation into three steps, allowing the model to better plan the generation: (1) the model is instructed to generate several ideas (*titles*) for a story that can demonstrate the use of the given words. (2) the model is asked to select the most interesting idea and write an *outline* for such a story. (3) the model is asked to turn the outline into a *final story* of a given length, with each selected word used at least three times.

Examples First In this method, we first prompt the model to generate *example sentences* for each selected word, before constructing the full story. This extra step is designed to help the model conceptualize the context and possible uses of the words, making their integration into the story more natural and coherent.

2.2.2 Lexical constraint enforcement

Once the story has been generated, it is processed by a non-LLM verifier. The verifier performs standard text normalization, tokenization, and lemmatization, then it checks the vocabulary used in the story against the list of words known to the user (see App. F for details). A list of broken lexical constraints, i.e. words previously unknown to the user and not selected by SRS (i.e., $x_i \notin V \cup L$), is produced.

Then, we use the following three strategies to fix lexical constraints:

Rewrite (base) We simply list the unknown words to the model and prompt it to replace them with simpler alternatives. This process is repeated five times.[2] The full prompt is given in Appx. B.

Rewrite Highlighted In this strategy, instead of giving the LLM a list of words to replace, the model is asked to simplify words that are highlighted in the text. The story is presented to the model with unknown words surrounded by asterisks (Markdown-style bold text).

Get Synonyms then Rewrite The model is first prompted to list synonyms for the listed words unknown to the user. In a second step, it is requested to rewrite the story using the synonyms provided.

3 Experiments

3.1 Experimental setup

In order to evaluate[3] the proposed story generation techniques, we conduct experiments on three languages: English, Polish, and Chinese.

Vocabulary levels For each language, we compiled a vocabulary list with the assigned language proficiency level. For English, we used the word lists provided by CEFR-J (Negishi et al., 2013; Tono, 2016), a Japanese standard for English language proficiency. For Chinese, the word lists from New HSK 3.0 standard were used (National Language Commission, 2021). Since no official word lists are available for Polish, we compiled them from a word frequency list computed on the Polish Wikipedia part of Leipzig Corpora Collection (Goldhahn et al., 2012). We assume the most frequent 1,500 words to be CEFR A1, 3,500 words for A2, 5,000 for B1, 7,500 for B2, 10,000 for C1, and all remaining words are assumed to be C2 words.

Study session setup We simulate a study session using the SRS-Stories system, assuming the user is familiar with all vocabulary at a given language level and aims to learn 10 random words from the next level.[4] Note that using random words is a challenging setup that does not exploit dependencies between words, such as being related to a single lesson topic. For each language level and method, the results are averaged over 200 stories.

LLMs used We generate the stories using the Llama 3.1 70B Instruct model (Grattafiori et al., 2024), an open-weight large language model that achieves high performance on many benchmarks and performed well in our preliminary experiments. Additionally, we also report the performance of other LLMs for comparison. We selected: the general-purpose LLM Qwen2.5 72B Instruct (Yang et al., 2025), the multilingual LLM Aya 23 35B (Aryabumi et al., 2024) and the reasoning LLM DeepSeek R1 70B (DeepSeek-AI et al., 2025). The stories were generated using the default parameters in the vLLM library (Kwon et al., 2023), except for the maximum number of tokens limited to 4096 and the temperature set to zero.

Baseline As a baseline method of generating text with lexical constraints, we use the HuggingFace (Wolf et al., 2020) implementation of Constrained Beam Search (CBS, Post and Vilar, 2018; Hu et al., 2019) to generate texts that specifically use given words L. Additionally, to limit the vocabulary only to words known by the user (i.e., use $V \cup L$ only), HuggingFace offers the "No Bad Words" strategy of masking unwanted tokens at the output of the language model and setting their generation probability to zero. This approach adapts to multi-token words, masking out unwanted word

[2] In our experiments, the number of replaced words flattens out after 5 iterations, and subsequent iterations are less effective.

[3] The implementation of our experiments is available here.

[4] We also experimented with a mixed setup, where some words are new (should have multiple occurrences in the story) and some are only reviewed (one occurrence is enough), but we found that teaching the user new words is much more challenging – see Appx. E.

continuations on-the-fly, based on tokens previously generated.

3.2 Metrics

The stories are evaluated by a set of automatic metrics: (1) assessment of different aspects of the generated text by an LLM evaluator, (2) count-based metrics measuring the fulfillment of lexical constraints. Additional metrics, including model-based ones, are provided in Appx. D.

LLM-based evaluation Previous works have shown great potential in using LLMs to imitate human evaluators in natural language generation (NLG) tasks (Hu et al., 2024; Kocmi and Federmann, 2023). Building upon this, we use an LLM to evaluate generated text. We prompt the model to return a score on a scale from 1 to 5 to assess: grammatical correctness (**Gram.**), text coherence (**Coh.**) and story interestingness (**Int.**). Following previous work (Hu et al., 2024; Kartáč et al., 2025), we prompt the model to first list the errors found in the assessed text, and then provide a numerical quality assessment. The exact prompts are provided in Appx. C.

As the evaluator, we use Qwen2.5 72B Instruct (Yang et al., 2025), which officially supports all languages used in our experiments. It also obtained the best results among open-weight models in a multilingual NLG evaluation presented by Chang et al. (2025).

Lexical constraints checks To evaluate the fulfillment of lexical constraints in generated stories, we compute several count-based metrics:

- **#** L – the average number of occurrences of each word to learn. The story should incorporate new vocabulary multiple times, providing various examples of word usage to reinforce learning and improve vocabulary retention.
- $\#|L| \geq 1$ – the percentage of words to learn that are actually introduced in the story.
- **Len.** – the length of a generated story (number of words for English and Polish, number of characters for Chinese).
- **OOV** – the percentage of out-of-vocabulary/unknown words ($x_i \notin V \cup L$). Ideally, the generated story should only use vocabulary that is known to the user. However, this is sometimes difficult to achieve, especially on lower language levels.

4 Results

4.1 Story generation strategy comparison

The results of the different story generation techniques are presented in Table 1 for English B1 (lower intermediate) language proficiency level.[5]

CBS is the lowest-performing technique in terms of grammaticality, coherence and interestingness. This is confirmed by our manual analysis: stories generated by CBS sometimes list the new words out of context and the stories are not fluent, with frequent grammatical errors. They also often change the plot in completely unexpected ways, for example by suddenly changing the protagonist of the story.

Comparing the proposed story generation strategies, *Simple prompting* achieved the highest coherence, the second-best grammaticality, and the highest interestingness – ex aequo with *Planning*. The *Planning* strategy used new words more frequently and produced slightly longer stories, though at the cost of worse grammatical accuracy. The *Example first* method yielded the highest grammatical correctness score but the lowest coherence and interestingness, exposing the learner least frequently to the target vocabulary.

With the base *Rewrite* approach to removing unknown words, all story generation methods achieved the best grammatical correctness scores, the highest number of occurrences of the studied vocabulary $\#L$, and the lowest percentage of OOV words. More advanced rewriting approaches led to higher interestingness scores across all story generation strategies and improved coherence in some variants, resulting in longer stories that, however, contained more words unfamiliar to the user.

Statistical analysis We also performed a statistical analysis of the English results, averaged across all language levels, using standard two-sample unpaired *t*-tests. The detailed results are provided in App. H.

Across all three quality aspects measured by the LLM-based metric, the differences between the baseline CBS and any of the proposed strategies were statistically significant, with low p-values ($p < 0.001$).

For grammaticality, the differences between story generation methods were significant; however, the differences among the lexical constraint

[5]Note that the story generation at the lower levels is more challenging due to the stronger lexical constraints.

Method	LLM Scoring Gram.	Coh.	Int.	Average # L	Len.	Percent of OOV	# $\|L\| \geq 1$
Constrained Beam Search	3.22	3.82	3.92	1.46	538.9	6.71%	94.90%
Simple prompting	**4.05**	4.22	4.01	**1.92**	578.12	**0.64%**	**95.40%**
+ Get Synonyms then Rewrite	3.92	4.21	4.12	1.91	599.85	5.03%	90.40%
+ Rewrite Highlighted	3.96	**4.30**	**4.16**	1.61	608.58	3.86%	81.10%
Examples first	**4.08**	**4.11**	3.94	**1.42**	542.09	**0.46%**	**95.25%**
+ Get Synonyms then Rewrite	3.96	4.08	**4.06**	1.33	586.20	4.80%	92.15%
+ Rewrite Highlighted	3.94	**4.11**	3.99	1.21	576.83	4.08%	83.85%
Planning	**3.98**	**4.17**	4.04	**2.68**	609.67	**0.01%**	**96.35%**
+ Get Synonyms then Rewrite	3.85	4.16	**4.16**	2.45	645.57	0.05%	90.80%
+ Rewrite Highlighted	3.81	4.16	4.11	2.39	645.48	0.04%	83.75%

Table 1: The results of different story generation strategies for English at B1 CEFR (lower intermediate) level. The metrics are defined in Sec. 3.2.

Language	Method	LLM Scoring Gram.	Coh.	Int.	Average # L	Len.	Percent of OOV	# $\|L\| \geq 1$
Polish	Simple prompting	3.53	4.01	3.89	**2.06**	406.36	**0.34%**	**88.55%**
	+ Get Synonyms then Rewrite	3.69	4.05	**3.95**	1.57	410.98	2.32%	65.85%
	+ Rewrite Highlighted	**3.81**	**4.08**	3.92	1.52	408.36	3.45%	60.47%
Chinese	Simple prompting	**3.99**	4.06	3.61	0.68	804.41	**1.72%**	18.93%
	+ Get Synonyms then Rewrite	3.75	**4.09**	**3.79**	3.05	844.57	7.29%	**79.63%**
	+ Rewrite Highlighted	3.79	**4.09**	3.73	**3.13**	856.88	6.23%	78.24%
English	Simple prompting	**4.04**	4.27	4.06	**1.82**	584.58	**0.52%**	**95.22%**
	+ Get Synonyms then Rewrite	3.95	4.27	**4.15**	1.75	603.91	4.00%	88.73%
	+ Rewrite Highlighted	3.97	**4.31**	4.14	1.18	618.64	3.11%	59.75%

Table 2: The results of story generation for different languages, averaged over three language proficiency levels: lower intermediate (CEFR B1, HSK 3), upper intermediate (CEFR B2, HSK 4) and lower advanced (CEFR C1, HSK 5). The metrics are defined in Sec. 3.2.

enforcement strategies were not statistically significant.

For coherence, the differences between stories produced by *Simple prompting* under different rewriting strategies were again not significant, although any variant of *Simple prompting* was statistically better than the other generation methods. *Examples first* was statistically significantly worse than *Planning*.

Regarding interestingness, *Simple prompting* with basic rewriting was significantly worse than when using more advanced rewriting strategies, but it was still significantly better than the other story generation methods. Again, *Examples first* was statistically significantly worse than *Planning*.

4.2 Multilingual evaluation

Given the good results of *Simple Prompting* and the higher computational cost of the other two generation strategies (more iterations with LLM), we focused the multilingual experiments on *Simple Prompting* only. The results for Polish, Chinese and English, averaged over three language proficiency levels from lower intermediate to lower advanced, are shown in Table 2.

The results for English averaged over different language levels are quite similar to the B1-level results in Sec. 4.1, with high LLM-based scores and *Rewrite* performing best at including new vocabulary into generated stories. The number of out-of-vocabulary words is lower, which is expected given that higher language levels use a wider base vocabulary.

The presented methods are also very successful in generating SRS-Stories in Chinese. The number of out-of-vocabulary words is relatively small and the new words are mentioned more than the required three times in the text. The LLM assessment of grammatical correctness is the highest for *Simple Prompting* with the simplest rewriting strategy (*Rewrite*). However, *Rewrite Highlighted* provides higher coherence, interestingness and is the most successful in frequently including the new words into the story. The high quality of the generated stories is also confirmed by our manual analysis, where we found the stories to be fluent and pro-

Model	Method	LLM Scoring Gram.	Coh.	Int.	Average # L	Len.	Percent of OOV	# $\lvert L\rvert \geq 1$
Llama 3.1 70B Instruct	Simple prompting	**4.04**	4.27	4.06	**1.82**	584.58	**0.52%**	**95.22%**
	+ Get Synonyms then Rewrite	3.95	4.27	**4.15**	1.75	603.91	4.00%	88.73%
	+ Rewrite Highlighted	3.97	**4.31**	4.14	1.18	618.64	3.11%	59.75%
Qwen 2.5 72B Instruct	Simple prompting	**4.09**	**4.76**	4.40	2.24	698.99	**3.78%**	74.60%
	+ Get Synonyms then Rewrite	3.95	4.61	**4.41**	**2.53**	721.60	6.99%	**83.55%**
	+ Rewrite Highlighted	4.03	4.65	4.39	1.40	720.89	6.29%	44.72%
DeepSeek R1 70B	Simple prompting	**4.03**	**4.57**	**4.42**	**1.83**	692.78	**5.52%**	**77.72%**
	+ Get Synonyms then Rewrite	3.58	4.12	4.14	1.52	698.32	6.83%	65.07%
	+ Rewrite Highlighted	3.86	4.27	4.26	1.12	764.82	9.66%	46.58%
Aya 23 35B	Simple prompting	4.05	4.37	4.02	0.60	314.06	**6.42%**	33.52%
	+ Get Synonyms then Rewrite	4.08	4.70	4.39	**1.49**	652.74	8.92%	**68.95%**
	+ Rewrite Highlighted	**4.21**	**4.84**	**4.49**	1.25	661.06	9.44%	52.82%

Table 3: The comparison of capabilities of different LLMs for English story generation, averaged over B1, B2 and C1 CEFR levels. The metrics are defined in Sec. 3.2.

viding meaningful word usage examples in correct Chinese grammatical structures.

For Polish, the stories generated by *Simple Prompting* are shorter than for the other languages. *Rewrite Highlighted* was the method that achieved the highest scores for grammaticality, coherence, and the second-best interestingness, but also a higher percentage of out-of-vocabulary words. Simple *Rewrite* was much more efficient in this respect, while also maximising the occurrence of new words.

4.3 Results with different language models

Table 3 shows the comparison story generation capabilities of different LLMs. Among the investigated models, Llama 3.1 was the most successful in eliminating words unknown to the user from the generated text. It also was the most efficient in including all new vocabulary into the story.

Qwen 2.5 achieved high scores on LLM-based metrics, but these may be biased since Qwen was also used as the evaluator (Li et al., 2024). DeepSeek achieves the best results for *Rewrite*, and using more advanced rewriting techniques seem to only harm the performance. Aya achieved the highest scores on grammatical correctness, coherence and interestingness, but was ineffective in including the new words into the story.

4.4 Human evaluation

We a conducted human evaluation experiment, comparing our proposed *Simple Prompting* strategy to the CBS baseline on 50 English stories generated by each method. To assess the multilingual capabilities of our approach, we also evaluated 50 Chinese and 50 Polish stories. The evaluation included five aspects on a 5-point Likert scale: grammatical correctness (**Gram.**), coherence (**Coh.**), interestingness (**Int.**), use of learned words in context (**Use**), and overall quality (**Over.**). See App. I for annotation guidelines.

We recruited 20 annotators through the Prolific platform who had previously studied the story language as a second language and had reached fluency. This background ensures reliable judgments of grammatical correctness while also providing sensitivity to the language-learning context. Each annotator evaluated 10 stories, plus one additional story that served as an attention check (see App. G).

The results are presented in Tab. 4. To compare our method with the baseline, we computed standard unpaired t-tests with 5% significance level. Whereas stories generated by CBS are of similar coherence and interestingness, the stories generated by the proposed approach are significantly more grammatically correct ($p = 0.0151$) and significantly better illustrate the use of studied words ($p < 0.0001$). Both these aspects are crucial in language learning applications.

Interestingly, the stories generated in Chinese are assessed by human annotators even better than those for English, achieving very high grammatical correctness, coherence, and scoring high on exemplary word usages. The overall quality of Polish stories is assessed as higher than for English, with similar quality of word usage, coherence, and grammatical correctness.

Validation of LLM-based metrics We run LLM-based evaluation metrics on the same stories that were assessed by human annotators and use the

Method	Gram.	Coh.	Int.	Use	Over.
English (CBS)	3.56	3.96	3.23	1.24	3.24
English (Ours)	**4.08**	3.92	2.94	**3.88**	3.08
Polish (Ours)	3.78	3.92	3.50	3.80	3.66
Chinese (Ours)	4.68	4.40	3.66	4.40	4.16

Table 4: The results of human evaluation of stories generated by Llama 3.1 70B with Constrained Beam Search (CBS) and *Simple Prompting* (Ours). Evaluation aspects are defined in Sec. 4.4. Statistically significant differences for English are in bold.

	Pearson's r			Spearman's ρ		
	Gram.	Coh.	Int.	Gram.	Coh.	Int.
Eng.	0.557	0.507	0.456	0.383	0.280	0.465
Ch.	0.281	0.426	0.349	0.298	0.326	0.336
Pol.	0.212	-0.149	0.070	0.218	-0.158	0.092

Table 5: Correlation between human annotators and LLM-based metrics.

scores for Grammaticality, Coherence, and Interestingness to compute Pearson and Spearman correlations between the human and LLM judgments.

The results are presented in Table 5. The observed correlations with human judgments for English are moderate and comparable to those reported on story generation datasets for state-of-the-art LLM-based metrics (Kartáč et al., 2025). The correlations for Chinese and for Polish are lower, highlighting the challenges of using LLMs in multilingual contexts.

5 Summary

In this work, we present several methods for story generation for language learners with specific lexical constraints. The proposed story generation task can be coupled with a Spaced Repetition System (SRS) to teach the user new vocabulary via reading an enjoyable story, specially crafted for their language level and known vocabulary.

The experimental evaluation, conducted on three languages and various language levels, demonstrated that modern LLMs prompted with the proposed strategies are capable to generate fluent, grammatically correct stories that introduce new vocabulary in context. These results highlight the potential of SRS-Stories as emerging language learning technology.

Limitations

The generated stories may reflect social biases present in the LLM's pretraining data, influencing character roles or cultural assumptions. While the system aims to use new target words and avoid unknown vocabulary, some out-of-vocabulary (OOV) words may still appear, potentially affecting comprehension. Additionally, the narrative quality can vary, sometimes resulting in minor inconsistencies or unnatural phrasing. Future improvements could enhance bias mitigation and vocabulary control.

Acknowledgments

This work was supported by the European Research Council (Grant agreement No. 101039303, NG-NLG) and used resources of the LINDAT/CLARIAH-CZ Research Infrastructure (Czech Ministry of Education, Youth, and Sports project No. LM2018101).

References

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Aidan Gomez, Phil Blunsom, Marzieh Fadaee, and 2 others. 2024. Aya 23: Open weight releases to further multilingual progress. *Preprint*, arXiv:2405.15032.

Yasin Aslan. 2011. Teaching vocabulary effectively through flashcards. *International Journal of Arts & Sciences*, 4(11):347.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Jiayi Chang, Mingqi Gao, Xinyu Hu, and Xiaojun Wan. 2025. Exploring the multilingual nlg evaluation abilities of llm-based evaluators. *Preprint*, arXiv:2503.04360.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of

deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

D. Elmes. 2015. Anki. http://ankisrs.net.

Dirk Goldhahn, Thomas Eckart, Uwe Quasthoff, and 1 others. 2012. Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages. In *LREC*, volume 29, pages 31–43.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

J. Edward Hu, Huda Khayrallah, Ryan Culkin, Patrick Xia, Tongfei Chen, Matt Post, and Benjamin Van Durme. 2019. Improved lexically constrained decoding for translation and monolingual rewriting. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 839–850, Minneapolis, Minnesota. Association for Computational Linguistics.

Xinyu Hu, Li Lin, Mingqi Gao, Xunjian Yin, and Xiaojun Wan. 2024. Themis: A reference-free NLG evaluation language model with flexibility and interpretability. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15924–15951, Miami, Florida, USA. Association for Computational Linguistics.

Ivan Kartáč, Mateusz Lango, and Ondřej Dušek. 2025. OpeNLGauge: An Explainable Metric for NLG Evaluation with Open-Weights LLMs. In *Proceedings of the 18th International Natural Language Generation Conference*, Hanoi, Vietnam. Association for Computational Linguistics.

Witold Kieraś and Marcin Woliński. 2017. Morfeusz 2 – analizator i generator fleksyjny dla języka polskiego. *Język Polski*, XCVII(1):75–83.

Tom Kocmi and Christian Federmann. 2023. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Sebastian Leitner. 1972. So lernt man lernen: Der weg zum erfolg [learning to learn: The road to success]. *Freiburg: Herder*.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. Llms-as-judges: A comprehensive survey on llm-based evaluation methods. *Preprint*, arXiv:2412.05579.

Katharine Pace Miles and Linnea C. Ehri. 2017. Learning to read words on flashcards: Effects of sentence contexts and word class in native and nonnative english-speaking kindergartners. *Early Childhood Research Quarterly*, 41:103–113.

Paul Nation and Karen Wang. 1999. Graded readers and vocabulary.

National Language Commission. 2021. Chinese proficiency grading standards for international chinese language education.

Masashi Negishi, Tomoko Takada, and Yukio Tono. 2013. A progress report on the development of the cefr-j. In *Exploring language frameworks: Proceedings of the ALTE Kraków Conference*, pages 135–163.

Matt Post and David Vilar. 2018. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1314–1324, New Orleans, Louisiana. Association for Computational Linguistics.

Siddharth Reddy, Igor Labutov, Siddhartha Banerjee, and Thorsten Joachims. 2016. Unbounded human learning: Optimal scheduling for spaced repetition. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1815–1824, New York, NY, USA. Association for Computing Machinery.

Yukio Tono. 2016. The cefr-j wordlist version 1.6.

Rob Waring and Misako Takaki. 2003. At what rate do learners learn and retain new vocabulary from reading a graded reader?

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System*

Demonstrations, pages 38–45, Online. Association for Computational Linguistics.

Piotr Woźniak and Edward Gorzelańczyk. 1994. Optimization of repetition spacing in the practice of learning. *Acta neurobiologiae experimentalis*, 54(1):59–62.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

A Frequently Asked Question

1. *In the paper, V is the set of words believed to be known by the user. However, the experiments are performed with the words assumed to be known according to CEFR/HSK lists which is different from the words really known to the user.*

   While we conducted experiments in simulated environments where an artificial student was assumed to know all words at a given CEFR/HSK level, our system retrieves the list of known words directly from the SRS. The SRS records all previously learned vocabulary and dynamically separates it into two categories: words that require review (because they may have been forgotten or are close to being forgotten) and words that are retained (still known). Ideally, learners should use the SRS from the very beginning of their language learning journey.

   If a learner does not start with the SRS and already possesses some vocabulary knowledge, most systems either estimate their vocabulary size basing on the language level (as in our experiments) or provide word lists aligned with textbooks to help the user construct the list of known words on their own. To bring the most benefits, SRS should store and schedule reviews of all known vocabulary. The challenge of initializing the SRS with a learner's existing knowledge is not unique to our approach but is inherent to all SRS-based systems.

2. *How are verb tenses handled by SRS-Stories?*

   Lemmatization is applied for constraint verification i.e. if the model generates a story containing different form of a required word, the system will accept it. Nevertheless, the model is asked to provide a story with a given verb form. Posing hard requirements on specific word forms would be impractical especially for Polish, which features relatively complex morphology not just for verbs but also nouns and other parts of speech.

3. *How are phrases or idioms handled?*

   Our experiments assume simple vocabulary for simplicity, but there is little difference in learning/using a lexicalized meaning of a longer phrase. Our approach has no technical limitations in this regard and can be directly used for phrase/idiom learning without any modifications.

4. *How are function words handled in CEFR lists?*

   Common function words such as "a" and "the" are part of the A1 English level. Other function words are often associated with specific tenses or grammatical structures (especially in Chinese) and are therefore assigned to different language levels according to their difficulty. The problem of learning how to use "a" or "the" is typically not related to recalling their meanings, so these words cannot be effectively learned using standard SRS. However, readers can gain an intuitive understanding of how to use these words based on the context by reading our SRS-Stories, which naturally use these function words frequently. This is one of the advantages of using our SRS-Stories vs. traditional SRS.

B Prompts

Simple prompting

```
Write a story (500-750 words)
containing words: [WORDS TO LEARN]. Use
them in such a way that it is clear
from the context of this story what
they mean. Use each word at least three
times. Write only story and nothing
more.
```

Planning

```
I would like to learn some new words in
English. These words are: [WORDS TO
LEARN]. Suggest some possible ideas for
stories that can include all these
words. Write only titles of stories and
```

nothing more. Number stories with letters a, b, c... instead of numbers 1, 2, 3...

Pick the most interesting idea. Write only the idea title and nothing more. Do not explain why you picked this one.

Generate a plan in points for a story (500-750 words) based on the idea you have picked. It should be possible to generate a full story, based on this plan, which will contain all the words that I want to learn: [WORDS TO LEARN].

Step by step generate a full story (500-750 words) based on this plan. Remember to use words which I want to learn: [WORDS TO LEARN]. Use them is such a way that it is clear from the context of this story what they mean. Use each word at least three times. Write only story and nothing more.

Examples first

I will give you a list of words. For each of these words provide a sentence that will clearly illustrates what this word means without any prior knowledge of this word. Words for which you should generate sentences: [WORDS TO LEARN]

Write a story (500-750 words) containing sentences that you have generated. If possible, try to use all of them. If not, remember to include all of the words [WORDS TO LEARN] in the story anyway. Make your story consistent and interesting to read. Write only a story and nothing more.

Rewrite

Rewrite this story without these words: [UNKNOWN WORDS]. Use simpler alternatives of these words. Change only this words and nothing else. Do not remove these words: [WORDS TO LEARN]. Write only the story and nothing more. Do not inform me that here is a story.

Rewrite Highlighted

I will provide you a story with some words marked with an asterisks (*) before and after a word, in other words these words are bolded. Change all of these words to simpler alternatives. Change only these words and nothing else. Do not remove these words: [WORDS TO LEARN]. Story with marked words: [HIGHLIGHTED STORY]. Write only the story and nothing more. Do not inform me that here is a story.

Get Synonyms then Rewrite

In the following text there are several words taken in brackets (). Could you list for each of them a list of synonyms? [HIGHLIGHTED STORY]

Re-write the above story by replacing the words in brackets with their synonyms. Don't do any other changes. Write only a story and nothing more.

C LLM-based evaluation prompts

Grading grammar

I will provide you a story. List all the grammatical errors in this story. Next give this story a grade from 1 to 5 (1 being the worst possible grade and 5 being the best possible grade). When grading a story focus on grammatical errors, do not take into account any other aspects. Story: [STORY]

Grading coherence

I will provide you a story. Analyse whether this story is coherent or not. Next give this story a grade from 1 to 5 (1 being the worst possible grade and 5 being the best possible grade). When grading a story, focus on its coherence, do not take into account grammatical errors.

Grading interestingness

I will provide you a story. Analyse whether this story is interesting to read or not. Next give this story a grade from 1 to 5 (1 being the worst possible grade and 5 being the best possible grade). When grading a story

```
focus on whether the story is
interesting or not, do not take into
account grammatical errors.
```

D Detailed results

Detailed results for English are in Table 6, for Chinese in Table 7, and for Polish in Table 8.

In addition to metrics shown in the main paper, we also report the following:

- L **to len.** – the ratio of new words to story length, provides an estimation of how often the new words occur in the story.
- Perplexity (**Perp.**) – we use the Qwen2.5 7B model (Yang et al., 2025) to compute the perplexity of the generated text.
- We use the BERT model (Devlin et al., 2019) finetuned on Corpus of Linguistic Acceptability (**COLA**; Warstadt et al., 2019) to assess whether each sentence in the story is grammatically correct. We report the percentage of sentences assessed as correct.
- Next-sentence prediction (**NSP**) is a binary classification task that predicts whether two sentences appear consecutively in the text. We use the NSP head of the BERT model (Devlin et al., 2019) to assess the coherence of the story.

Due to the suitability of models, the latter three metrics are computed for English only.

E Results for mixed setup

We also conducted preliminary experiments with Simple Prompting in a setup where the story would not only contain new words that would be used multiple times in meaningful contexts, but also review words that could only be mentioned. In our experimental setup, we used 3 random words as new words and 7 random words as words to review. The results are presented in Tab. 9.

F Details on lexical constraints verification

To check the lexical constraints, the Polish text was preprocessed using simple tokenisation with whitespaces and lemmatisation with Morfeusz (Kieraś and Woliński, 2017). Next, all possible conjugations/declinations of the lemmas were extracted from Morfeusz and checked against the list of permitted vocabulary. For the English stories, tokenisation and lemmatisation were performed using the nltk library (`WordNetLemmatizer`) (Bird and Loper, 2004). No tokenisation or lemmatisation was applied to Chinese. Therefore, to assess out-of-vocabulary metrics, we counted the number of Chinese characters that could not be mapped to the vocabulary lists for the given level. Before processing the text, all punctuation was removed and the text was converted to lowercase for all languages.

G Human evaluation details

The task of annotating eleven stories was estimated to take 45 minutes. The annotators were paid £6 for task completion (equivalent to £8 per hour), which is within the 'fair' pay range on the Prolific platform. One of the stories was used as an attention check: the order of the paragraphs had been switched (coherence), many obvious grammatical errors had been introduced manually (correctness) and none of the studied words were used (word use). Annotators who assessed these aspects as higher than 3 (the middle of the score range) were rejected from the study. The average assessment of the story used for the attention check is provided in Table 10.

We used four screening filters provided by Prolific for annotators selection: 1) Annotators had to be fluent in the language of the story (English, Polish or Chinese). 2) The language of the story had to be a language other than the annotators' first language. 3) The task approval rate had to be above 95%. 4) The number of tasks performed on the platform had to be above 50. The annotation was performed via Google Forms. The study was approved by the Ethics Committee of the authors' institution.

H Statistical analysis

The results of two-sample unpaired T-test performed for English stories are presented in Tables 11, 12, and 13.

I Annotation guidelines

This task concerns the evaluation of short stories (texts) intended as learning material for students of Chinese at different levels. Each story is accompanied by a list of words that can be learned by reading it.

Your task is to carefully read 11 such stories and their word lists in order to answer the following

level	Method	LLM Scoring Gram.	Coh.	Int.	Average # L	Len.	L to len.	Percent of OOV	# $\|L\| \geq 1$	Average NSP	Perp.	COLA
B1	Simple prompting	4.05	4.22	4.01	1.92	578.12	29.69	0.64%	95.40%	0.96	5.16	0.98
B1	+ Get Synonyms then R.	3.92	4.21	4.12	1.91	599.85	30.86	5.03%	90.40%	0.96	5.55	0.98
B1	+ Rewrite Highlighted	3.96	**4.30**	**4.16**	1.61	608.58	37.17	3.86%	81.10%	**0.97**	**5.01**	0.98
B1	Planning	3.98	4.17	4.04	**2.68**	609.67	**22.34**	0.52%	**96.35%**	0.95	5.68	0.98
B1	+ Get Synonyms then R.	3.85	4.16	**4.16**	2.45	645.57	25.95	4.99%	90.80%	0.95	6.13	0.98
B1	+ Rewrite Highlighted	3.81	4.16	4.11	2.39	645.48	26.62	4.36%	83.75%	0.95	5.60	0.98
B1	Examples first	**4.08**	4.11	3.94	1.42	542.09	37.66	**0.46%**	95.25%	0.96	5.40	**0.99**
B1	+ Get Synonyms then R.	3.96	4.08	4.06	1.33	586.20	43.65	4.80%	92.15%	0.96	5.82	**0.99**
B1	+ Rewrite Highlighted	3.94	4.11	3.99	1.21	576.83	46.78	4.08%	83.85%	0.96	5.34	0.98
B2	Simple prompting	4.04	**4.33**	4.08	1.87	580.32	30.57	**0.46%**	95.55%	0.96	**4.91**	0.98
B2	+ Get Synonyms then R.	3.96	4.31	**4.16**	1.70	605.04	35.44	3.64%	89.15%	**0.97**	5.15	0.98
B2	+ Rewrite Highlighted	4.00	4.32	4.12	1.22	600.93	48.84	2.59%	62.30%	**0.97**	4.93	0.98
B2	Planning	3.99	4.18	4.07	**2.33**	616.34	**26.39**	0.56%	94.45%	0.95	5.47	0.98
B2	+ Get Synonyms then R.	3.90	4.19	4.11	2.05	643.70	31.08	3.40%	85.25%	0.96	5.63	0.98
B2	+ Rewrite Highlighted	3.90	4.16	4.07	1.79	647.58	35.86	3.30%	67.00%	0.95	5.30	0.98
B2	Examples first	**4.10**	4.18	4.03	1.33	556.63	41.42	0.47%	**95.80%**	0.96	5.14	**0.99**
B2	+ Get Synonyms then R.	3.98	4.17	4.10	1.26	588.54	46.53	3.54%	90.90%	0.96	5.25	**0.99**
B2	+ Rewrite Highlighted	3.95	4.20	4.03	0.97	589.09	59.56	3.03%	66.00%	0.96	4.99	0.98
C1	Simple prompting	4.04	4.27	4.09	1.68	595.29	35.27	0.46%	94.70%	0.96	5.11	**0.98**
C1	+ Get Synonyms then R.	3.98	**4.29**	**4.16**	1.63	606.83	36.96	3.34%	86.65%	**0.97**	5.16	**0.98**
C1	+ Rewrite Highlighted	3.94	**4.29**	4.13	0.70	646.41	92.02	2.87%	35.85%	**0.97**	**4.92**	**0.98**
C1	Planning	3.97	4.23	4.05	**2.10**	624.02	29.57	**0.45%**	93.70%	0.95	5.45	**0.98**
C1	+ Get Synonyms then R.	3.90	4.21	4.12	1.84	651.84	35.04	3.24%	82.35%	0.95	5.62	**0.98**
C1	+ Rewrite Highlighted	3.94	4.13	4.08	1.14	655.00	57.48	3.12%	45.55%	0.95	5.38	**0.98**
C1	Examples first	**4.08**	4.16	4.04	1.21	574.54	47.19	0.46%	**95.20%**	0.95	5.11	**0.98**
C1	+ Get Synonyms then R.	3.98	4.18	4.12	1.07	590.78	55.01	3.47%	87.05%	0.96	5.26	**0.98**
C1	+ Rewrite Highlighted	3.99	4.16	4.06	0.54	598.36	**109.69**	2.97%	42.80%	0.96	5.00	**0.98**
AVG	Simple prompting	4.04	4.27	4.06	1.82	584.58	31.84	0.52%	95.22%	0.96	5.06	0.98
AVG	+ Get Synonyms then R.	3.95	4.27	**4.15**	1.75	603.91	34.42	4.00%	88.73%	**0.97**	5.28	0.98
AVG	+ Rewrite Highlighted	3.97	**4.31**	4.14	1.18	618.64	59.34	3.11%	59.75%	**0.97**	**4.95**	0.98
AVG	Planning	3.98	4.20	4.05	**2.37**	616.68	**26.10**	0.51%	94.83%	0.95	5.53	0.98
AVG	+ Get Synonyms then R.	3.88	4.19	4.13	2.12	647.03	30.69	3.88%	86.13%	0.95	5.79	0.98
AVG	+ Rewrite Highlighted	3.88	4.15	4.09	1.77	649.35	39.99	3.59%	65.43%	0.95	5.43	0.98
AVG	Examples first	**4.09**	4.15	4.00	1.32	557.75	42.09	**0.46%**	**95.42%**	0.95	5.22	**0.99**
AVG	+ Get Synonyms then R.	3.97	4.14	4.09	1.22	588.51	48.39	3.94%	90.03%	0.96	5.44	**0.99**
AVG	+ Rewrite Highlighted	3.96	4.16	4.03	0.91	588.09	72.01	3.36%	64.22%	0.96	5.11	0.98

Table 6: Detailed results for English, Llama-3.1-70B-Instruct

level	method	LLM Scoring Gram.	Coh.	Int.	Average # L	Len.	L to len.	Percent of OOV	# $\|L\| \geq 1$
HSK 3	Simple prompting	**3.94**	4.09	3.54	0.94	809.93	86.55	2.77%	27.38%
HSK 3	+ Get Synonyms then Rewrite	3.63	**4.15**	**3.77**	3.13	849.03	26.65	**10.11%**	**81.38%**
HSK 3	+ Rewrite Highlighted	3.68	4.09	3.70	**3.43**	881.21	**24.78**	8.23%	81.26%
HSK 4	Simple prompting	**3.98**	4.08	3.65	0.77	804.31	**104.57**	**1.73%**	21.49%
HSK 4	+ Get Synonyms then Rewrite	3.71	4.05	**3.78**	2.60	840.04	32.33	5.83%	77.41%
HSK 4	+ Rewrite Highlighted	3.79	**4.09**	3.70	**2.87**	846.47	28.84	5.34%	**78.96%**
HSK 5	Simple prompting	**4.05**	4.01	3.65	0.33	798.98	244.53	**0.65%**	7.91%
HSK 5	+ Get Synonyms then Rewrite	3.92	**4.08**	**3.83**	**3.41**	844.63	**24.74**	5.94%	**80.10%**
HSK 5	+ Rewrite Highlighted	3.88	**4.08**	3.78	3.10	842.95	26.74	5.11%	74.50%
AVG	Simple prompting	**3.99**	4.06	3.61	0.68	804.41	**145.22**	**1.72%**	18.93%
AVG	+ Get Synonyms then Rewrite	3.75	**4.09**	**3.79**	3.05	844.57	27.90	7.29%	**79.63%**
AVG	+ Rewrite Highlighted	3.79	**4.09**	3.73	**3.13**	856.88	26.79	6.23%	78.24%

Table 7: Detailed results for Chinese, Llama-3.1-70B-Instruct

level	method	LLM Scoring Gram.	Coh.	Int.	Average # L	Len.	L to len.	Percent of OOV	# $\|L\| \geq 1$
B1	Simple prompting	3.54	4.08	3.89	**2.04**	390.99	**19.14**	**0.48%**	**89.60%**
B1	+ Get Synonyms then Rewrite	3.69	**4.13**	**3.96**	1.57	388.01	24.69	3.31%	69.45%
B1	+ Rewrite Highlighted	**3.78**	4.11	3.93	1.61	386.63	24.07	5.20%	69.50%
B2	Simple prompting	3.59	4.03	3.89	**1.95**	402.29	**20.59**	**0.22%**	**87.80%**
B2	+ Get Synonyms then Rewrite	3.64	4.07	**3.98**	1.50	396.50	26.45	1.91%	69.80%
B2	+ Rewrite Highlighted	**3.83**	**4.11**	3.93	1.46	403.66	27.74	2.68%	64.85%
C1	Simple prompting	3.46	3.94	3.90	**2.17**	425.81	**19.63**	**0.33%**	**88.25%**
C1	+ Get Synonyms then Rewrite	3.75	3.95	**3.92**	1.64	448.43	26.50	1.73%	58.30%
C1	+ Rewrite Highlighted	**3.83**	**4.02**	3.88	1.49	434.78	28.09	2.47%	47.05%
AVG	Simple prompting	3.53	4.01	3.89	**2.06**	406.36	**19.79**	**0.34%**	**88.55%**
AVG	+ Get Synonyms then Rewrite	3.69	4.05	**3.95**	1.57	410.98	25.88	2.32%	65.85%
AVG	+ Rewrite Highlighted	**3.81**	**4.08**	3.92	1.52	408.36	26.63	3.45%	60.47%

Table 8: Detailed results for Polish, Llama-3.1-70B-Instruct

level	method	LLM Scoring Gram.	Coh.	Int.	New+recall Avg. # L	% of # $\|L\| \geq 1$	New Avg. # L	% of # $\|L\| \geq 1$	Recall Avg. # L	% of # $\|L\| \geq 1$	Avg. Len.	% of OOV
B1	Simple pr.	4.04	4.34	4.00	4.47	93.20%	3.33	29.60%	1.38	63.60%	572.76	4.39%
B2	Simple pr.	3.94	4.22	4.04	2.87	94.40%	3.48	29.60%	1.19	64.80%	548.54	2.93%
C1	Simple pr.	4.02	4.32	4.14	3.31	91.40%	3.05	28.40%	1.17	63.00%	563.48	3.04%
AVG	Simple pr.	4.00	4.29	4.06	3.55	93.00%	3.29	29.20%	1.25	63.80%	561.59	3.45%

Table 9: English, Llama-3.1-70B-Instruct, Words to be learned divided into new and recall sets

Method	Gram.	Coh.	Int.	Use	Over.
English	1.90	1.90	1.70	1.10	1.30
Polish	2.60	3.60	3.20	1.00	2.80
Chinese	2.00	2.20	1.40	1.00	1.80

Table 10: The results of human evaluation of the story with artificially introduced errors, which served as an attention check for the annotators.

questions:

- How fluent (i.e. grammatically correct) is the story?
- How coherent is the story (i.e. is the plot logical and consistent)?
- How interesting/enjoyable to read is the story (given the language learning context)?
- How well does the story illustrate the use of words from the list ? Are the words from the word list used in the story, and are they used correctly? Does their use help to guess their meaning?
- What is your overall evaluation of the generated story?

You should answer each question by selecting a score from 1 (lowest) to 5 (highest).

Warning: The stories are generated by an Artificial Intelligence algorithm and are entirely fictional.

J Examples of generated stories

J.1 English

An example of a story generated at B2 level for words *carriage, core, environmentally, imagine, relaxing, reluctant, sure, timely, tuition, unexpectedly* is provided below.

In the small town, in the heart of the countryside, a beautiful old *carriage* stood proudly in the town square. The *carriage*, with its carefully carved wooden panels and soft seats, had been a part of the town's history for generations. It was said that the *carriage* had once belonged to a wealthy family who had used it to travel to and from their grand estate on the outskirts of town.

As a child, she had always been fascinated by the *carriage*. She would often think about herself riding in it, feeling like a princess on her way to a grand ball. But as she grew up, her fascination with the *carriage* never went away. In fact, it became stronger. She began to see the *carriage* as a symbol of the town's rich history and a reminder of the importance of preserving the past at its *core*.

As she entered her adult years, she became increasingly passionate about living in a way that

	Simple p.	+ Get Syn.	+ Rewrite.	Example first	+ Get Syn.	+ Rewrite.	Planning	+ Get Syn.	+ Rewrite.
CBS	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Simple prompting	-	0.000	0.000	0.009	0.000	0.000	0.001	0.000	0.000
+ Get Synonyms then Rewrite	-	-	0.153	0.000	0.125	0.278	0.037	0.000	0.000
+ Rewrite Highlighted	-	-	-	0.000	0.460	0.352	0.220	0.000	0.000
Example first	-	-	-	-	0.000	0.000	0.000	0.000	0.000
+ Get Synonyms then Rewrite	-	-	-	-	-	0.314	0.246	0.000	0.000
+ Rewrite Highlighted	-	-	-	-	-	-	0.133	0.000	0.000
Planning	-	-	-	-	-	-	-	0.000	0.000
+ Get Synonyms then Rewrite	-	-	-	-	-	-	-	-	0.467
+ Rewrite Highlighted	-	-	-	-	-	-	-	-	-

Table 11: P-values of two-sample unpaired T-tests between the results of *Grammaticality* obtained by different generation methods for English, averaged across all levels.

	Simple p.	+ Get Syn.	+ Rewrite.	Example first	+ Get Syn.	+ Rewrite.	Planning	+ Get Syn.	+ Rewrite.
CBS	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Simple prompting	-	0.424	0.126	0.000	0.000	0.000	0.001	0.000	0.000
+ Get Synonyms then Rewrite	-	-	0.092	0.000	0.000	0.000	0.001	0.000	0.000
+ Rewrite Highlighted	-	-	-	0.000	0.000	0.000	0.000	0.000	0.000
Example first	-	-	-	-	0.350	0.408	0.022	0.051	0.469
+ Get Synonyms then Rewrite	-	-	-	-	-	0.266	0.008	0.021	0.376
+ Rewrite Highlighted	-	-	-	-	-	-	0.035	0.076	0.376
Planning	-	-	-	-	-	-	-	0.358	0.016
+ Get Synonyms then Rewrite	-	-	-	-	-	-	-	-	0.040
+ Rewrite Highlighted	-	-	-	-	-	-	-	-	-

Table 12: P-values of two-sample unpaired T-tests between the results of *Coherence* obtained by different generation methods for English, averaged across all levels.

	Simple p.	+ Get Syn.	+ Rewrite.	Example first	+ Get Syn.	+ Rewrite.	Planning	+ Get Syn.	+ Rewrite.
CBS	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Simple prompting	-	0.000	0.000	0.000	0.061	0.018	0.255	0.000	0.068
+ Get Synonyms then Rewrite	-	-	0.412	0.000	0.005	0.000	0.000	0.205	0.001
+ Rewrite Highlighted	-	-	-	0.000	0.013	0.000	0.000	0.294	0.006
Example first	-	-	-	-	0.000	0.083	0.001	0.000	0.000
+ Get Synonyms then Rewrite	-	-	-	-	-	0.000	0.011	0.030	0.428
+ Rewrite Highlighted	-	-	-	-	-	-	0.047	0.000	0.000
Planning	-	-	-	-	-	-	-	0.000	0.010
+ Get Synonyms then Rewrite	-	-	-	-	-	-	-	-	0.013
+ Rewrite Highlighted	-	-	-	-	-	-	-	-	-

Table 13: P-values of two-sample unpaired T-tests between the results of *Interestingness* obtained by different generation methods for English, averaged across all levels.

helped the *environment*. She believed that it was essential to live in a way that was good for the planet and to reduce the harm we cause. She began to think of ways to restore the old *carriage* in a way that would be good for the *environment*. She wanted to make *sure* that the *carriage* would be a place where people could go to *relax* and enjoy the peaceful atmosphere of the countryside.

Her friends and family were initially uncertain to support her idea. They thought it would be a waste of time and money to restore the old *carriage*. But she was certain that it was a project worth doing. She spent many hours researching and planning, determined to find a way to make it work in a *timely* manner.

Finally, after months of hard work, her vision began to take shape. She had found a team of skilled workers who were willing to work with her to restore the *carriage* using methods that were good for the *environment*. The project was not without its challenges, but she was determined to see it through.

As the work progressed, the people of the town began to take notice. They would stop and watch as she and her team worked hard to bring the *carriage* back to life. Slowly but *surely*, the *carriage* began to transform before their eyes. The wooden panels were sanded and polished, the seats were made new again, and the metal parts were restored to their former shine.

When the work was complete, the people of the town gathered to admire the finished product. The *carriage* was more beautiful than anyone had ever thought possible. It was a true masterpiece, and a reminder of the importance of living in a way that is good for the *environment*. The *carriage* was a place where people could go to *relax* and enjoy the peaceful atmosphere of the countryside, and it was a *timely* reminder of the importance of preserving the past.

The *carriage* was officially opened to the public again, and it quickly became a popular place to visit. People would come from all over to ride in the beautifully restored *carriage*, enjoying the *relaxing* atmosphere of the countryside. The *carriage* also became a reminder of the importance of living in a way that is good for the *environment*, and it was a place where people could go to learn about living *environmentally*.

As the *carriage*'s popularity grew, so did her reputation as a skilled worker and *environmentalist*. She began to receive requests from all over the country to work on similar projects. She was happy to agree, using her skills to help preserve the past and promote *environmentally* friendly living.

Years later, her work on the *carriage* was recognized by a prestigious university, which offered her a full scholarship to study *environmental* science. She was very happy and accepted the offer without hesitation. She went on to become a leading expert in her field, using her knowledge to help protect the *environment* and preserve the past.

The old *carriage* remained a beloved part of the town, a reminder of her hard work and dedication. It stood as a reminder of the power of imagination and perseverance, and the impact that one person can have on their community. And every time she sat in the *carriage*, she was reminded of the importance of living in a way that is good for the *environment*, and she was *sure* that she had made a difference.

But what she did not expect was that the *carriage* would also bring her a surprise. One day, as she was sitting in the *carriage*, she met a man who would become her future husband. He was a historian, and he had come to the town to research the town's history. They struck up a conversation, and she showed him around the town, including the *carriage*. As they sat together, they became close, and the *carriage* became a special part of their love story, and it was a *timely* and *relaxing* experience for both of them."

J.2 Polish

An example of a story generated at B2 level for words *dojeżdżać* (to commute), *monitorować* (to monitor), *mordować* (to murder), *nieznaczny* (insignificant, minor), *nowatorski* (innovative), *ocaleć* (to be saved), *piotrowy* (Peter's), *rejonowy* (district's), *wynająć* (to rent, to hire), *zwężać* (to narrow) is provided below.

W małym miasteczku, położonym w sercu Polski, mieszkał młody i pełen energii Piotr. Był on znanym i cenionym w swoim *rejonie* za swoje *nowatorskie* podejście do pracy. Piotr miał zwyczaj *dojeżdżać* każdego dnia do pracy rowerem, co pozwalało mu na *monitorowanie* sytuacji w mieście i jego okolicach.

Jednego dnia, Piotr otrzymał informację o tajemniczym *morderstwie*, które miało miejsce w jednej z okolicznych wsi. Ofiarą był lokalny przedsiębiorca, znany ze złego postępowania. Piotr postanowił wyjaśnić sprawę i *wynająć* prywatnego detektywa, aby pomógł mu w śledztwie.

W trakcie śledztwa, Piotr i detektyw odkryli, że *morderstwo* nie było przypadkowe. Był to jeden z serii *morderstw*, które miały miejsce w okolicy w ciągu ostatnich kilku tygodni. Sprawca, który okazał się być bardzo niebezpiecznym człowiekiem, *mordował* swoje ofiary w okrutny sposób, pozostawiając za sobą tylko *nieznaczne* ślady.

Piotr i detektyw postanowili *monitorować* teren, aby znaleźć jakiekolwiek wskazówki, które mogłyby doprowadzić ich do sprawcy. W trakcie obserwacji, zauważyli, że w okolicy pojawia się coraz więcej patroli policyjnych, co mogło wskazywać na to, że policja również jest blisko rozwiązania sprawy.

Jednakże, Piotr i detektyw byli pewni, że są bliżej rozwiązania sprawy niż policja. Ich *nowatorskie* podejście do śledztwa pozwoliło im na odkrycie kilku ważnych wskazówek, które mogły doprowadzić ich do sprawcy.

W końcu, po kilku dniach intensywnych poszukiwań, Piotr i detektyw znaleźli się w *rejonie*, gdzie prawdopodobnie ukrywał się sprawca. Był to opuszczony budynek na obrzeżach miasta, który wyglądał jak idealne miejsce dla niebezpiecznego człowieka.

Gdy weszli do budynku, Piotr i detektyw zostali zaatakowani przez sprawcę. Był to moment, w którym musieli *zwężać* jego pole widzenia, aby uniknąć ataku. Jednakże, dzięki swojej szybkości i szybkiej reakcji, udało im się *ocaleć* i powstrzymać sprawcę.

Po aresztowaniu sprawcy, Piotr i detektyw zostali uznani za bohaterów. Ich *nowatorskie* podejście do śledztwa i umiejętność *monitorowania* sytuacji pozwoliły im na rozwiązanie sprawy, która wydawała się być niemożliwa do rozwiązania. Piotr powrócił do swojej pracy, ale teraz był znany nie tylko w swoim *rejonie*, ale w całym kraju.

J.3 Chinese

An example of a story generated at HSK4 level for words: 单调(monotonous), 团结(to unite), 干什么(what to do?), 必要(necessary, indispensable), 打折(discount), 把(grammar particle), 旅客(traveler), 条件(condition, circumstance), 电影院(cinema), 难道(could it be?) is provided below.

李明觉得自己的生活很单调，没有什么意思。每天都是重复的工作和生活，没有任何变化。他常常问自己："我干什么呢？难道这就是我的生活吗？"

小张对李明说："我们应该团结起来，做一些有意义的事情。我们可以一起做志愿者，帮助那些需要帮助的人。"李明觉得这是一个好主意。他们开始一起做志愿者，帮助老人、病人和需要帮助的人。

他们还组织了一个社区活动，让所有的居民参加。活动中，他们有游戏、音乐和电影。李明觉得自己的生活突然变得有趣起来。小张说："我们团结起来，做一些有意义的事情，生活就会变得不再单调。"

但是，有些人不理解他们的行为。他们觉得李明和小张是在浪费时间。他们说："你们干什么呢？这不是必要的工作啊。"李明和小张回答说："我们是在帮助别人，我们是在做有意义的事情。这是非常必要的。"

有一天，李明和小张决定去电影院看电影。他们买了票，准备进去的时候，发现电影院正在打折。他们觉得很幸运，能够以便宜的价格看电影。他们进去后，发现电影院里有很多旅客。他们觉得很奇怪，问电影院的工作人员："为什么这里有这么多旅客？"工作人员回答说："我们正在举办一个活动，很多人都来参加。"

李明和小张觉得很有趣。他们决定参加活动，和旅客们一起看电影。他们发现，活动的条件很好，电影也很精彩。他们觉得自己的生活变得更加丰富。小张说："我们把生活变得更加丰富了。"

李明突然想起了自己的问题："难道这就是我的生活吗？"他觉得自己已经找到了答案。他决定继续做志愿者，帮助别人，做有意义的事情。他觉得自己的生活已经变得不再单调。小张说："我们团结起来，生活就会变得更好。"