Automatically Suggesting Diverse Example Sentences for L2 Japanese Learners Using Pre-Trained Language Models

Enrico Benedetti[1*], **Akiko Aizawa**[2], **and Florian Boudin**[2,3]

[1]University of Bologna, Italy
[2]National Institute of Informatics, Japan
[3]JFLI, CNRS, Nantes University, France
enrico.benedetti5@studio.unibo.it, aizawa@nii.ac.jp
florian.boudin@univ-nantes.fr

Abstract

Providing example sentences that are diverse and aligned with learners' proficiency levels is essential for fostering effective language acquisition. This study examines the use of Pre-trained Language Models (PLMs) to produce example sentences targeting L2 Japanese learners. We utilize PLMs in two ways: as quality scoring components in a retrieval system that draws from a newly curated corpus of Japanese sentences, and as direct sentence generators using zero-shot learning. We evaluate the quality of sentences by considering multiple aspects such as difficulty, diversity, and naturalness, with a panel of raters consisting of learners of Japanese, native speakers – and GPT-4. Our findings suggest that there is inherent disagreement among participants on the ratings of sentence qualities, except for difficulty. Despite that, the retrieval approach was preferred by all evaluators, especially for beginner and advanced target proficiency, while the generative approaches received lower scores on average. Even so, our experiments highlight the potential for using PLMs to enhance the adaptability of sentence suggestion systems and therefore improve the language learning journey.

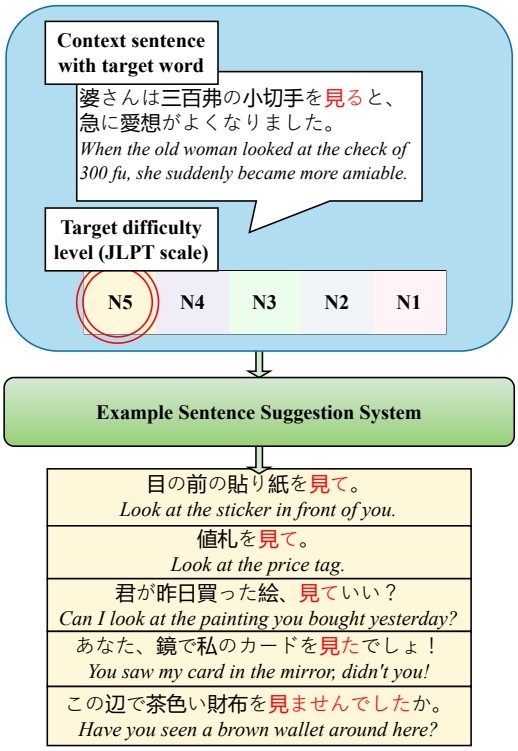

Figure 1: Task overview. Given a word in context and a difficulty level, the system will suggest diverse and level-appropriate examples. In this instance, the target is *miru*, to see.

1 Introduction

The term second language acquisition (or L2 acquisition) refers to the process of learning a second language by those who already know a first one. While children have a natural predisposition for acquiring languages, the degree of success among L2 learners varies greatly, as it is usually harder in adult life, requiring a combination of conscious effort, motivation, support from teachers and adequate materials (Fromkin et al., 2013).

Online dictionaries are usually the first resource towards which learners turn to in order to understand an unknown word or expression via definitions and example sentences. However, producing high-quality learning material requires effort and expert knowledge. Because of that, researchers have explored automated techniques for selecting and generating examples to aid professionals like lexicographers or teachers, as well as non-experts like language learners (Kilgarriff et al., 2008; Ward, 2017; Pilán et al., 2013a).

Pre-trained Language Models (PLMs) have been shown to be effective for many NLP tasks (Wang et al., 2023). The main motivation for this work is to investigate whether PLMs can be leveraged to propose sentences that are understandable and diverse to help L2 learners be exposed to a broad range of uses for the target words they are inter-

[*]Research conducted during internship at NII, Japan.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 114–131

ested in (e.g. an unknown word encountered while reading), since examples contribute to improving vocabulary knowledge (Baicheng, 2009).

In this study, we focus on Japanese, as an increasing number of people are interested in achieving a certain level of proficiency, be it for study, work, culture or other reasons (Nakamachi et al., 2022). While there is substantial work on obtaining high-quality text from corpora or generative models, as discussed in Section 2, to the best of our knowledge, there are few studies simultaneously addressing the Japanese example sentence suggestion task, and developments in Natural Language Processing (NLP) such as the emergence of PLMs. The existing work mostly focuses on functional expressions (Liu et al., 2018a,b; Liu and Matsumoto, 2016; Shortt, 2021) or exercises (Andersson and Picazo-Sanchez, 2023).

Our contributions are summarized as follows:

1. We develop a retrieval-based approach to select example sentences from a corpus, by combining different PLM modules and NLP techniques for scoring sentence quality according to four criteria: difficulty, sense similarity, syntactic and lexical diversity.
2. We build WJTSentDiL, a corpus of sentences from different web sources, annotated with Japanese Language Proficiency Test[1] (JLPT) labels.
3. We evaluate the quality of selected example sentences for specific target words by comparing the retrieval approach to two generative PLM baselines, employing native speakers and learners, alongside GPT-4 (OpenAI, 2023). We present the insights obtained from the investigation.

The main repository for this work can be found here: NihongoExamplePLM.

2 Related Work

In the following we discuss the related work, namely retrieving and generating example sentences, and estimating sentence difficulty.

Example selection Similarly to Tolmachev and Kurohashi (2017), we seek to provide high-quality and diverse example Japanese sentences. They propose a thorough retrieval approach based on quality and diversity scoring using a Determinantal Point Process, and carry out an evaluation with L2 learners and a teacher. Our work differs from theirs in that we focus on selecting sentences for sense similarity given a target word in context, instead of many possible senses for a word in isolation. Furthermore, we evaluate more aspects of the systems, in particular their capacity to adapt to learner proficiency levels. We also employ a language model in the evaluation.

Many other works deal with the task of example sentence selection from a corpus, focusing on dictionary examples for English, Japanese and Swedish (Kilgarriff et al., 2008; de Melo and Weikum, 2009; Hazelbeck and Saito, 2009; Pilán et al., 2013b). Additionally, Shinnou and Sasaki (2008), Kathuria and Shirai (2012) and Cheng et al. (2018) leverage parallel corpora to extract disambiguated sentences, while we limit our experiments to the monolingual setting.

Example generation There is a lot of research on controllable text generation approaches (Zhang et al., 2023a). Possible generation targets are definitions for a given term (Zhang et al., 2023b; Gardner et al., 2022), as well as example sentences. When it comes to example generation, researchers have shown that generated sentences can improve performance in Word Sense Disambiguation tasks in a supervised (Barba et al., 2021) or unsupervised way (He and Yiu, 2022). Focusing on L2 learners, Harvill et al. (2023) consider lexical complexity and sentence length to generate example sentences of controllable difficulty. In our case, we opt not to rely on fixed sense inventories, primarily due to the scarcity of available sense-tagged corpora. However, we believe that assigning dictionary definitions to words could prove beneficial to learners.

Sentence difficulty estimation Determining the level of difficulty of text is a key challenge in educational NLP, as vocabulary and grammatical structure interact in a complex way (Collins-Thompson, 2014). To estimate the difficulty of Japanese sentences, Nakamachi et al. (2022) show that a BERT-based classifier (Devlin et al., 2019) trained on labeled examples can achieve good performance, surpassing existing readability metrics[2] and approaches based on word frequencies. Liu and Matsumoto (2017) focus on estimating Japanese text difficulty for learners with pre-existing knowledge of Chinese characters. In that case, the main source of difficulty is not vocabulary, but grammar and

[1] More details on the JLPT website and Section 4.1.1.

[2] https://jreadability.net/sys/en

functional expressions. In our work, due to lacking training data from official JLPT material, we train a similar classifier to Nakamachi et al. (2022).

3 Task: Example Sentence Suggestion

We define the L2 contextualized example suggestion task as:

$$M(w, s_0, d) = \{s_1, s_2, \ldots, s_i, \ldots, s_K\} \quad (1)$$

Given a target word w, a context sentence s_0 and a target difficulty level d, we want to obtain a list of K good example sentences from a model M.

To expand more on what makes a good example, Kilgarriff et al. (2008) suggest that such examples should represent typical usage, be informative and understandable to learners. Building upon the discussion presented by Tolmachev et al. (2022), we aim to obtain multiple examples with diverse syntactic patterns since learners preferred them.

4 Methodology

4.1 Retrieval method

We design a retrieval model that, given a query, will select candidate sentences containing a target word from a corpus and present them to the learner (for more details on the corpus, see Section 5.1). Candidate sentences are ranked by how closely they match the target difficulty level and the semantic similarity of the target word in both the suggested and context sentences. Finally, the model selects a subset of sentences considering the total diversity of the list. In summary, we devise a model to quantify for a sentence s_i:

1. how adequate s_i is with respect to the target difficulty level d (Sec. 4.1.1).
2. if s_i contains the target word w and it is used in the same sense as the target word of the context sentence (Sec. 4.1.2).
3. the diversity of $\{s_0, s_1, s_2, \ldots, s_i, \ldots, s_K\}$ on vocabulary and syntax (Sec. 4.1.3).

4.1.1 Quality: difficulty

The Japanese Language Proficiency Test (JLPT) has a proficiency scale similar to the Common European Framework of Reference for Languages (CEFR). The JLPT levels are, from easier to harder: N5, N4, N3, N2 and N1. Our classifier will therefore assign a JLPT level d_i to input sentences. Then, it will be mapped to a difficulty score between 1 and 0. We formulate this score as

$$max\big(0, 1 - \text{penalty}_{\text{diff}} * (d - d_i)\big) \quad (2)$$

where d and d_i are the target difficulty level and difficulty label of the sentence i. We manually set the coefficient $\text{penalty}_{\text{diff}}$ to 0.2. We increase the coefficient to 0.4 on sentences deemed harder than the target level because L2 learners might benefit more from easier sentences in case of discrepancies.

4.1.2 Quality: sense similarity

Pilehvar and Camacho-Collados (2019) propose Words in Context (WiC), a different declination of Word Sense Disambiguation. WiC is a binary classification task: given a target word and two contexts, the model has to predict whether the word is used with the same meaning. Since we also tackle this problem in our case, we turn to MirrorWiC, an unsupervised fine-tuning method for contextualized word sense embeddings (Liu et al., 2021a). We fine-tune a PLM with MirrorWiC and use the resulting model to extract a vector representation for the target words in context. Then, we assign a sense similarity score based on cosine similarity between s_0, the context sentence, and s_i.

4.1.3 Diversity: syntactical and lexical

Inspired by the way Tolmachev and Kurohashi (2017) measured syntax diversity, we opt for a simpler approach, supported by other works on syntax similarity (Chen et al., 2023a; Kanagawa and Okadome, 2016).

We compute dependency trees of two sentences and partially generalize their labels, then apply a Label-based Tree Kernel Similarity method, FastKASSIM, to obtain a diversity score (Chen et al., 2023a; Moschitti, 2006; Boghrati et al., 2018). More in detail, we compute the parse trees and the number of shared subtrees of a pair of sentences. The latter is normalized with the square root of the product of the number of subtrees for each sentence (Chen et al., 2023a). For the syntactic diversity of a list of sentences, we take the average of pairwise scores.

For lexical diversity, we simply compute the average percentage of unique 1-2-3-4-grams in a sentence list.

Finally, we obtain a combined diversity score by equally weighting the lexical and syntax scores.

4.1.4 Ranking and Greedy Selection

As the number of candidates can be very high, we greedily select K final sentences. First, we sort the candidate sentences in terms of difficulty and sense scores, having equal weights as we consid-

ered the qualities equally important for this experiment. Then, within a window, we iteratively add the sentence which achieves the highest diversity score, until the list is complete. We set a window of only 50 candidates in the preliminary experiments. Otherwise, queries would take a long time due to having to re-compute similarity scores for every partial list.

4.2 PLM generation method

Considering the PLM baselines, we prompt them with the query, expressed in English. We share the prompt used in Appendix C. As initial experiments revealed that complying with the query in zero-shot manner was quite difficult, we prompt the PLMs multiple times, concatenate the outputs and exclude duplicates and sentences without the target word, until we get the required number of sentences. In the majority of cases, twice was enough. We set the generation temperature parameter to 1.0 for all PLMs; additionally, for LLM-jp, we add a repetition penalty of 5.0.

5 Experimental setup

5.1 Dataset: WJTSentDiL Corpus

We present WJTSentDiL,[3] a corpus of **W**ikipedia, **J**pWaC and **T**atoeba **Sent**ences with **Di**fficulty **L**evel. It is built by merging together three public corpora (described below) and performing additional filtering to remove spurious sentences. Additionally, our difficulty classifier adds JLPT levels to each sentence.

- Tatoeba is a platform where users can share sentences and translations. We select only Japanese sentences and fix errors where entries are made from multiple sentences.
- JpWaC (Sangawa et al., 2010) is a curated corpus of sentences automatically collected from Japanese web domains. We include subsets L0 to L4 of the corpus.
- Wikipedia is a free online encyclopedia. We process raw article text from the Japanese part of the website, more specifically the "jawiki dump" from December 2023.

We use spaCy[4] and Ginza[5] to split raw text into sentences, tokenize them, and assign part-of-speech (POS) tags. To keep well-formed sentences, we apply filters following heuristics similar to Kilgarriff et al. (2008) and Sangawa et al. (2010). Namely, we keep sentences that:

- have a length between 5 and 50 tokens.
- have less than 20% punctuation or numerals.
- do not contain tokens from the Latin, Cyrillic and Arabic scripts.
- end in a predicate and punctuation, or particles such as よ, ね.
- are not duplicates.

Wikipedia sentences are what makes up most of the corpus. They are on average longer and contain more *kanji*, Chinese characters, compared to the other sources. We show statistics in Table 1.

Corpus	Sentences	Tokens	Kanji (%)	Ratio (%)
JpWaC	152 751	13.01	27.31	1.2
Tatoeba	245 793	11.07	26.75	1.9
Wikipedia	12 306 416	26.39	36.67	96.9
WJTSentDiL	12 704 960	25.93	36.35	100

Table 1: Statistics of WJTSentDiL by source. "Tokens" is the average token count."Kanji" reports the proportion between Chinese characters and the rest.

5.2 Retrieval method details

5.2.1 Inverted index

The retrieval model uses an inverted index, mapping words to sentences they appear in. The keys are lemmas or "dictionary forms" of words and compound nouns. The candidate sentences are retrieved using the index by lemmatizing the target word. For example, the target word "たべた" (past form of *to eat*) is lemmatized as "食べる+た" (*to eat* + past tense auxiliary verb).

5.2.2 Difficulty classifier

The JLPT difficulty classifier is a BERT model pretrained on texts in the Japanese language,[6] that we fine-tuned on 5,000 sentences from Japanese language learning websites.[7] Their labels are assigned based on HTML metadata specific to each website. For more details on the training and evaluation of the classifier, see Appendix A. Its performance is very good (84% accuracy) on in-distribution data (i.e. the validation split), but it worsens on a different test set composed of official JLPT past exam sentences (38% accuracy). Our hypothesis is that

[3] The corpus is available on HuggingFace.
[4] Repository for spaCy, version 3.7.2
[5] Repository for ginza, version 5.1.3, 'ja-ginza' model.
[6] tohoku-nlp/bert-base-japanese-v3
[7] nihongokyoshi-net.com, jlptsensei.com. Due to license limitations, we can not share the sentences, but the model is available on HuggingFace.

the latter test set contains very long sentences composed of many relative clauses, which are very different from the sentences used for training.

5.2.3 Sense embeddings

We use MirrorWiC (Liu et al., 2021a) to fine-tune multiple baseline PLMs with 10,000 sentences randomly chosen from our corpus. To guide model selection, we look at their performance on two WiC tasks, XL-WiC (Raganato et al., 2020) and AM2iCo (Liu et al., 2021b). MirrorWiC fine-tuning shows a small improvement on both tasks for BERT-base-japanese, over the same base model and a Japanese Sentence Transformer.[8]

To obtain the embeddings, we average the last 4 layers of the embedding model, and across the sub-tokens that make up the target word, following Liu et al. (2021a).

6 Evaluation

6.1 Goals of the evaluation

We outline the core research questions that guide our investigation.

Q1: The capabilities of LLMs such as GPT-4 in rating text have been explored (Chen et al., 2023b). Therefore, can GPT-4 evaluate the quality of Japanese sentences from the perspective of L2 learners, and how do its assessments compare to those given by humans?

Q2: How do the automated quality metrics we used to guide the development of the retrieval approach compare with human judgment?

Q3: How good are PLMs at following instructions for this complex task?

Q4: Is text retrieved from a corpus (assumed to be human-authored) preferred to generated text?

Q5: What do humans think of their output?

We try to answer those questions by asking volunteer L2 learners and Japanese native speakers to manually rate and rank systems outputs.

6.2 Selected baselines

The systems we consider are the retrieval approach (Section 4.1), LLM-jp, a Japanese PLM,[9] and GPT-3.5. Specifically, throughout the paper, when mentioning GPT-3.5 we mean GPT-3.5-turbo-0613, while GPT-4 is GPT-4-0125-preview.

[8] sonoisa/sentence-bert-base-ja

[9] llm-jp/llm-jp-13b-instruct-full-jaster-dolly-oasst-v1.0

6.3 Evaluation data preparation

We build a set of target words from those used in the human evaluation of Tolmachev and Kurohashi (2017) and also add words from a work in WSD by Okumura et al. (2011). The former study involved 14 target words, and the latter 50, sharing one word, resulting in a total of 63. We randomly divided them into 53 for validation and experimental use, and 10 for testing and human evaluation, but ensuring a test set composition of 3 nouns, 4 verbs, 2 adjectives, and 1 adverb.

In addition, for every target word, we obtain a context sentence by randomly selecting sentences from *yourei* and *gogo*,[10] websites which provide a search engine for snippets of text content.

6.4 Human evaluation guidelines

We consider as a query the input for the task (Equation 1), namely the selected word for human evaluation, along with their associated context sentence and target level. In this experiment we target levels N1, N3 and N5. The system outputs are randomly ordered and presented with the query, forming an "annotation block". Each baseline provides $K = 5$ sentences. This results in 30 blocks (10 queries $\times$ 3 levels), and 150 sentences for each system (30 blocks $\times$ 5 output sentences).

We ask evaluators to rate:

1. **Difficulty level**, by rating the difficulty of each sentence on the JLPT scale. This is to see how closely systems match the target difficulty.
2. **Sense similarity**, by evaluating whether the usage of the target word in each sentence aligns with its sense in the original context. This is to see whether the proposed sentences retain the use of the word in a similar sense.
3. **Rejection**: sentences should be marked for rejection if they are deemed not useful (e.g. unnatural usage) or confusing (e.g., grammatical or segmentation errors).
4. **Syntactic Diversity**, by examining the variety in sentence structure and the different grammatical constructions used to incorporate the target word.
5. **System Ranking**: after rating each system's outputs, rank them from best to worst. The ranking should consider the overall utility for language learners at the target proficiency.

[10] https://yourei.jp, https://dictionary.goo.ne.jp

We demonstrated the task and explained the evaluation guidelines. The total participants are 5, of which 3 are native Japanese speakers and 2 are learners of proficiency N1-N2. Addionally, we include an annotation block example in Appendix B.

6.5 PLM evaluation protocol

We feed GPT-4 a modified version of the evaluation guidelines, the system outputs, and ask it to rate them. More details can be found in Appendix D. Empirically, we noticed that ratings for the same prompt sometimes were different, even when trying to reduce variability. So, we query GPT-4 three times, and also obtain its majority vote. We note that in some cases this could still result in an unclear rating.

7 Results and Discussion

In this section, we present the results from the evaluation. We try to address our research questions in three main parts: agreement between raters; systems comparison; comments and error analysis.

7.1 Q1: Agreement of ratings

The Intraclass correlation coefficient (ICC) is a widely used statistical measure for reliability, that reflects the degree of correlation and agreement between ratings (Koo and Li, 2016). The reason for choosing this metric is that it takes into account the magnitude of the differences between scores. For example, it is important that if a sentence is rated N1 by one person and N5 by another, it is seen as a larger disagreement than one rated N1 and N2.

We compute the metric with the pingouin library,[11] and we convert ratings from ordinal labels into numbers, mapping them in a scale where the relative distances are the same among labels. Following Hackl et al. (2023), who studied the reliability of GPT-4 in a similar experiment, we use a specific ICC setting based on a two-way mixed effect model. In short, ICC(3,1), according to the naming convention of Shrout and Fleiss (1979).

7.1.1 GPT-4 rating consistency

In Table 2, we report ICC values for the quality ratings across groups of raters. We include in this table only raters who compiled at least half of the blocks for each target level, in order to have a generalizable idea of the agreement.

[11] https://pingouin-stats.org/build/html/index.html

For GPT-4, despite setting its behavior to be nearly deterministic and obtaining ratings on the same day, we observed that the consistency of its ratings varies by type. The model shows excellent agreement in assessing JLPT levels and good consistency in rejecting sentences. However, its consistency is lower for other evaluation areas like sense similarity, syntax diversity, and model ranking. Using a mean combination of ratings improves consistency, but comes at the cost of more forward passes on the same long inputs. A way to further mitigate this is improving the prompt.

7.1.2 Agreement among groups

Focusing on human raters, it seems that agreement on qualities except difficulty level is quite low (Table 2). One reason for this could be that the guidelines for other metrics are too generic, which causes more variability in the ratings. However, we expected that language learners and native speakers may not have the same rating patterns. Additionally, since we required many ratings at once, there could be some additional effects at play, such as fatigue or bias from the order of annotation.

7.1.3 Pairwise agreement on ranking

To further investigate whether GPT-4 ranks similarly to humans, in Table 3 we report the pairwise agreement for the preferred system ranking from all annotators.

Inter-rater agreement between GPT-4 and humans is generally lower than those among humans of different groups. This suggests that humans, regardless of whether they are native speakers or not, have more similar ranking preferences compared to the AI models. However, there are also outliers, such as HN2, who has a way of ranking that shows no agreement with many other raters. This highlights the challenge in aligning AI evaluations with human preferences and confirms that, even among humans, there is significant disagreement on judging learning material suitability.

7.2 Q2-3-4: Quantitative analysis of ratings

After the agreement analysis, we discuss how raters evaluated the systems. For qualities other than difficulty and ranking preference, we report the main empirical findings in the following, and release additional figures in Appendix E.

7.2.1 Difficulty level ratings

Figure 2 shows the proportion of human-assigned JLPT difficulty labels for each system, grouped by

Rater group→	GPT-4 ($N = 3$)		Human ($N = 3$)		All ($N = 4$)	
Rated item↓	**ICC(3,1)**	**95% CI**	**ICC(3,1)**	**95% CI**	**ICC(3,1)**	**95% CI**
Level	0.941	[0.93, 0.95]	0.681	[0.63, 0.73]	0.673	[0.63, 0.72]
Sense	0.640	[0.59, 0.68]	0.258	[0.18, 0.33]	0.108	[0.06, 0.17]
Reject	0.861	[0.84, 0.88]	0.238	[0.18, 0.30]	0.244	[0.20, 0.30]
Syn. diversity	0.778	[0.70, 0.84]	0.214	[0.08, 0.36]	0.236	[0.13, 0.36]
Ranking	0.694	[0.60, 0.78]	0.218	[0.09, 0.36]	0.218	[0.12, 0.34]

Table 2: ICC estimates and their 95% confidence intervals (CI) for different groups. N indicates the number of raters in the group. In the last group, we consider the humans and the majority vote of GPT-4.

Rater↓→	GPT-4_{majority}	GPT-4_1	GPT-4_2	GPT-4_3	HL 1	HL 2	HN 1	HN 2	HN 3
GPT-4_{majority}	1	0.80*	0.78*	0.93*	0.37*	0.22*	0.37*	0.05	0.20
GPT-4_1	0.80*	1	0.55*	0.72*	0.33*	0.17	0.35*	0.02	0.11
GPT-4_2	0.78*	0.55*	1	0.82*	0.29*	0.17	0.45*	0.13	0.28*
GPT-4_3	0.93*	0.72*	0.82*	1	0.37*	0.21*	0.28*	-0.03	0.20
HL 1	0.37*	0.33*	0.29*	0.37*	1	0.29*	0.46*	0.13	0.68*
HL 2	0.22*	0.17	0.17	0.21*	0.29*	1	0.22*	0.14	0.47*
HN 1	0.37*	0.35*	0.45*	0.28*	0.46*	0.22*	1	0.30*	0.42*
HN 2	0.05	0.02	0.13	-0.03	0.13	0.14	0.30*	1	0.42*
HN 3	0.20	0.11	0.28*	0.20	0.68*	0.47*	0.42*	0.42*	1

Table 3: Pairwise agreement matrix of ICC(3,1) scores on **ranking preferences**. "HL" refers to a human learner, while "HN" to a human native speaker. *: *P*-value is less than .05.

System→	Retrieval				LLM-jp				GPT-3.5			
Rater↓, **Target**→	N1	N3	N5	Tot.	N1	N3	N5	Tot.	N1	N3	N5	Tot.
GPT-4_{majority}	**7**	**5**	**5**	**17**	2	2	2	6	1	3	3	7
HL 1†	**5**	**4**	–	**9**	0	0	–	0	0	2	–	2
HL 2	**4**	3	**6**	**13**	2	2	2	6	**4**	**4**	2	10
HN 1	**10**	**4**	**10**	**24**	0	2	0	2	0	**4**	0	4
HN 2	**7**	1	**8**	**16**	2	**5**	1	8	1	4	1	6
HN 3†	**7**	1	–	**8**	1	2	–	3	0	**6**	–	6

Table 4: Number of annotation blocks in which the considered baseline is rated first in overall quality, by target difficulty level. †: The participant mostly rated blocks with target level N1 and N3 only, because of time constraints.

target level. When considering how close the difficulty of proposed sentences is to the target level, our retrieval approach is markedly better for N1 and N5, while for N3, it produced a significant proportion of harder sentences. GPT-3.5 seems better for N3, but being so consistent is not always an advantage because it makes it difficult to adapt to user requirements, for example when requesting advanced sentences. LLM-jp also had issues following the prompt: repetitions, sentences without the target word, incoherent text.

7.2.2 Sense similarity ratings

When the raters indicated whether the target word in each sentence had a similar meaning as the one in the context, the vast majority classified the sense as being the same. The percentage of sentences rated as "not similar" was only about 2% for the retrieval, and 13% for the generative baselines. This shows that the systems generally succeed in producing examples with similar nuances.

7.2.3 Rejection ratings

According to our evaluation guidelines, unnatural sentences and those with confusing errors should be marked. On average, 8% of sentences suggested by the retrieval were rejected, while for LLM-jp it was 13%, and 16% for GPT-3.5.

Checking raters' comments confirmed that there were some segmentation errors in retrieval and generation baselines, such as sentences starting with punctuation, or with a fragment. It seems that generative models are more prone to errors, while the retrieved sentences are better in this aspect "by design". Still, careful text pre-processing and post-processing is needed as sentences with errors can

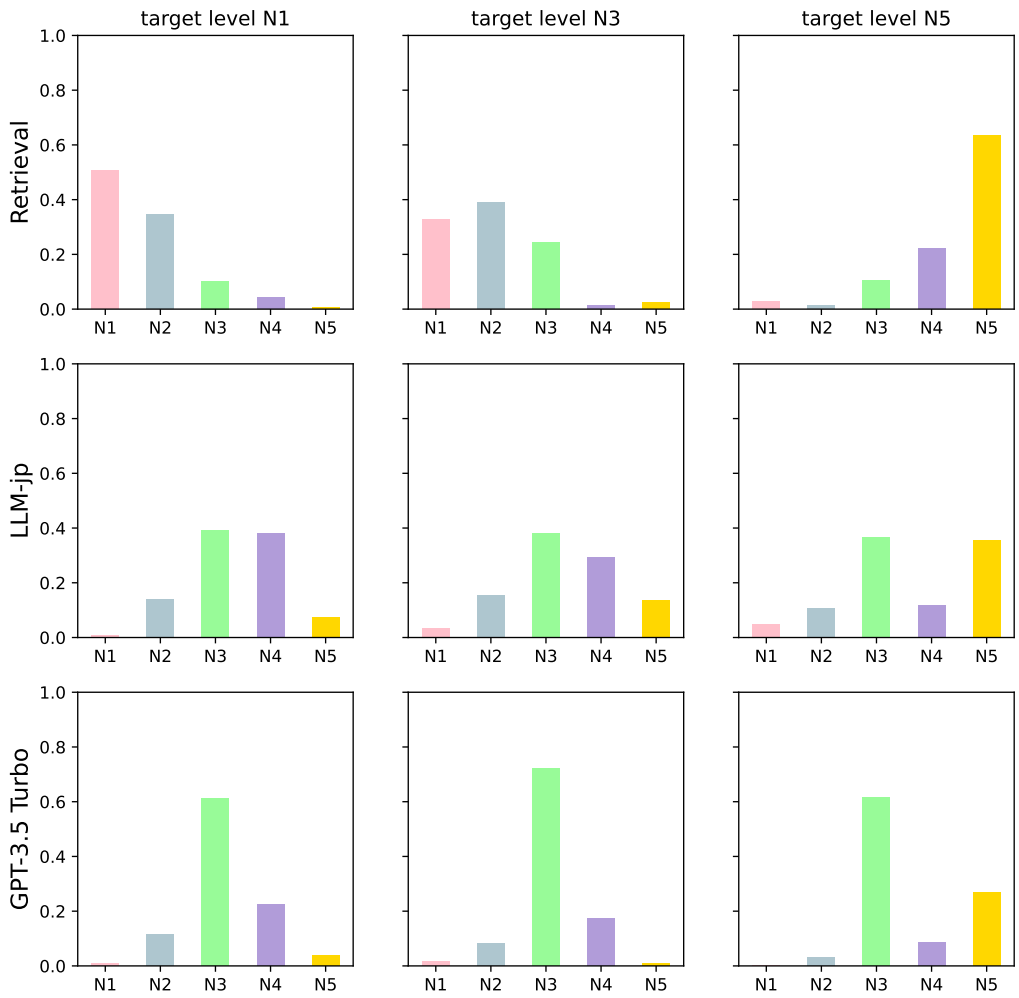

Figure 2: Evaluators' ratings on difficulty. Each row presents the proportions of JLPT labels assigned by humans for one system, across the three target difficulty levels set for the evaluation.

be confusing for beginner learners.

For a couple of concrete examples, the following sentence from the retrieval approach was rejected by some human participants because it sounded unnatural and too literary: この闘いは今日の場合では大概は容易ならぬ苦闘だからだ。 *"As for this fight, in today's situation, it is generally a difficult struggle"*.

Finally, the following was generated by LLM-jp and was rejected because of the presence of confusing characters and English words at the beginning: `favorite dish is sushi.1.`右手で持っていたスプーンを左手でも持てるようになったんだ。 *"The spoon that I was holding with the right hand, I became able to hold with the left hand as well"*.

7.2.4 Diversity ratings

Considering the syntax diversity of the list of sentences, the retrieval method earned the most "high" ratings across all target levels. GPT-3.5 received mostly "medium" votes, and LLM-jp got the lowest. The latter model often produced repetitive sentences, where only one or two words would differ between each generated sentence. This highlights another issue in zero-shot generation, i.e. that it is difficult to have both diversity and adherence to instructions.

7.2.5 System ranking ratings

Table 4 presents votes on system ranking by human participants and GPT-4. The sentence lists produced by the retrieval system are the best overall for all raters when considering the total vote count. Except for HL2 and HN3, the retrieval system is rated best in over 50% of cases. When considering target levels, it also markedly wins in suggesting lists for advanced and beginner target difficulty levels, while it is not rated best as much for the intermediate level. The sentences suggested by

the retrieval system for N3 are often on the more difficult end, as shown in Figure 2.

7.3 Q5: Qualitative analysis and participants comments

A native speaker commented on a target word in the evaluation (全然, *zenzen*). It is commonly used in negative statements, to mean "not at all" (Sawada, 2007). Using it in positive statements can be considered "slightly broken" in formal situations, but it was correct a hundred years ago, and it is used in today's slang. In that case, GPT-3.5 produced a similar sentence as the context in which the usage was "uncommon". Indeed, the context sentence was from an excerpt of a work published in 1938 by Osamu Dazai, a famous Japanese writer. This should prompt thinking about what actually makes a correct sentence. Language learners noted that many sentences contained one or two difficult kanji, encountered at higher proficiency levels, even though the overall sentence structure is more straightforward to understand. This happened mostly with the retrieval approach, which did not take word difficulty explicitly into account.

8 Conclusion

This paper outlines a methodology for suggesting example sentences to learners of Japanese. It is adaptable to other languages with minor adjustments. The baselines we consider highlight many possible roles of PLMs: assessing difficulty, encoding semantic representations, directly producing sentences and evaluating their quality, all of which could be investigated further on the basis on their applicability in AI-supported language learning and other fields in education technology.

From the feedback and data collected from the human evaluation, we can point out the potential for improving and combining these systems to balance their shortcomings, even though the retrieval methodology was considered to be the best in terms of diversity and adherence to difficulty level.

The challenge of evaluating generated text prompted us to explore a state-of-the-art LLM's ability in rating sentence quality. In our opinion, it is a promising direction because the model seems to be able to evaluate linguistic features of sentences. We found good agreement in rating text difficulty, but since each person could make different assessments, finding a way to take that variability into account could be useful for personalization.

It could be studied whether using word-level features can prevent unknown kanji from appearing in example sentences. Such features could be JLPT labels or the school grade level they are taught in. Another research challenge is estimating the real vocabulary known by the learner, modeling the process of second language acquisition (Settles et al., 2018; Cui and Sachan, 2023). Additionally, there is potential for suggestion and generation of material based on each learner's interests.

A direction to explore further is to experiment with more advanced LLM prompting strategies, such as chain of thought or reinforcement learning, to iteratively refine outputs for better adaptation to learners' preferences. A retrieval approach like ours could serve as a starting point.

Limitations

In our work, the retrieval approach scores sentences using mainly unsupervised approaches and PLMs.

The corpus we build is not as large as other corpora. In our comparisons, for LLMs we explored only basic prompting strategies without fine-tuning, wanting to investigate approaches in a setting without labeled data.

As for the evaluation, the number of volunteers who participated in the study was quite limited and the agreement values are not very high, indicating that the results are not generalizable to larger groups. Nevertheless, we believe that the feedback and guidelines could be valuable for future research. About half of them were foreign students, and their feedback was valuable. Unfortunately, due to lack of resources, none of the native speakers were language educators. Involving language teachers would be advisable. Additionally, comparing our baselines with the approach of Tolmachev and Kurohashi (2017) would have been insightful. However, due to the absence of a practical implementation and limited resources for human evaluation, we opted for PLM baselines.

Ethics Statement

Because of the training methods of base LLMs, sentences generated or retrieved using these approaches could reflect negative biases that could impact or influence negatively the model of language that is internalized by the learners. It poses an increased risk when there are not enough sources of information, or limited sharing of ideas and communication with other learners and native speakers

of the foreign language that can more effectively teach distinguishing polite and casual register and other aspects of pragmatics, other than just word usage.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 24K03231. We thank the LLM Research and Development Centre for LLMs, National Institute of Informatics, Japan, for the support.

We also wish to thank the volunteer participants in the evaluation experiments for their help, Dr. Arseny Tolmachev for the precious advice, and the anonymous reviewers for their feedback.

References

Tim Andersson and Pablo Picazo-Sanchez. 2023. Closing the gap: Automated distractor generation in japanese language testing. *Education Sciences*, 13(12).

Zhang Baicheng. 2009. Do example sentences work in direct vocabulary learning? *Issues in Educational Research*, 19.

Edoardo Barba, Luigi Procopio, Caterina Lacerra, Tommaso Pasini, and Roberto Navigli. 2021. Exemplification Modeling: Can You Give Me an Example, Please? In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 3779–3785, Montreal, Canada. International Joint Conferences on Artificial Intelligence Organization.

Reihane Boghrati, Joe Hoover, Kate M Johnson, Justin Garten, and Morteza Dehghani. 2018. Conversation level syntax similarity metric. *Behavior research methods*, 50(3):1055–1073.

Maximillian Chen, Caitlyn Chen, Xiao Yu, and Zhou Yu. 2023a. Fastkassim: A fast tree kernel-based syntactic similarity metric. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*.

Yi Chen, Rui Wang, Haiyun Jiang, Shuming Shi, and Ruifeng Xu. 2023b. Exploring the use of large language models for reference-free text quality evaluation: An empirical study. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, pages 361–374, Nusa Dua, Bali. Association for Computational Linguistics.

Shang-Chien Cheng, Jhih-Jie Chen, Chingyu Yang, and Jason Chang. 2018. LanguageNet: Learning to Find Sense Relevant Example Sentences. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 99–102, Santa Fe, New Mexico. Association for Computational Linguistics.

Kevyn Collins-Thompson. 2014. Computational assessment of text readability: A survey of current and future research. *ITL - International Journal of Applied Linguistics*, 165(2):97–135.

Peng Cui and Mrinmaya Sachan. 2023. Adaptive and personalized exercise generation for online language learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Volume 1: Long Papers*, pages 10184 – 10198, Stroudsburg, PA. Association for Computational Linguistics. 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023); Conference Location: Toronto, Canada; Conference Date: July 9-14, 2023.

Gerard de Melo and Gerhard Weikum. 2009. Extracting sense-disambiguated example sentences from parallel corpora. In *Proceedings of the 1st Workshop on Definition Extraction*, pages 40–46, Borovets, Bulgaria. Association for Computational Linguistics.

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.

V. Fromkin, R. Rodman, and N. Hyams. 2013. *An Introduction to Language*. Cengage Learning.

Noah Gardner, Hafiz Khan, and Chih-Cheng Hung. 2022. Definition modeling: literature review and dataset analysis. *Applied Computing and Intelligence*.

Veronika Hackl, Alexandra Elena Müller, Michael Granitzer, and Maximilian Sailer. 2023. Is GPT-4 a reliable rater? Evaluating consistency in GPT-4's text ratings. *Frontiers in Education*, 8.

John Harvill, Mark Hasegawa-Johnson, Hee Suk Yoon, Chang D. Yoo, and Eunseop Yoon. 2023. One-Shot Exemplification Modeling via Latent Sense Representations. In *Proceedings of the 8th Workshop on Representation Learning for NLP (RepL4NLP 2023)*, pages 303–314, Toronto, Canada. Association for Computational Linguistics.

Gregory Hazelbeck and Hiroaki Saito. 2009. A Corpus-based E-learning System for Japanese Vocabulary. *Journal of Natural Language Processing*, 16(4):3–27.

Xingwei He and Siu Ming Yiu. 2022. Controllable Dictionary Example Generation: Generating Example Sentences for Specific Targeted Audiences. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 610–627, Dublin, Ireland. Association for Computational Linguistics.

Eriko Kanagawa and Takeshi Okadome. 2016. Syntactic characteristics and similarities of japanese authors' writing styles: A kernel-based approach. In *2016 International Conference on Asian Language Processing (IALP)*, pages 59–62.

Pulkit Kathuria and Kiyoaki Shirai. 2012. Word Sense Disambiguation Based on Example Sentences in Dictionary and Automatically Acquired from Parallel Corpus. In *Advances in Natural Language Processing*, Lecture Notes in Computer Science, pages 210–221, Berlin, Heidelberg. Springer.

Adam Kilgarriff, Milos Husák, Katie McAdam, Michael Rundell, and P. Rychlý. 2008. Gdex: Automatically finding good dictionary examples in a corpus. In *Proceedings of the 13th EURALEX International Congress*, Barcelona, Spain. Institut Universitari de Linguistica Aplicada, Universitat Pompeu Fabra.

Terry K. Koo and Mae Y. Li. 2016. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, 15(2):155–163.

Jun Liu, Fei Cheng, Yiran Wang, Hiroyuki Shindo, and Yuji Matsumoto. 2018a. Automatic Error Correction on Japanese Functional Expressions Using Character-based Neural Machine Translation. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*, Hong Kong. Association for Computational Linguistics.

Jun Liu and Yuji Matsumoto. 2016. Simplification of Example Sentences for Learners of Japanese Functional Expressions. In *Proceedings of the 3rd Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA2016)*, pages 1–5, Osaka, Japan. The COLING 2016 Organizing Committee.

Jun Liu and Yuji Matsumoto. 2017. Sentence complexity estimation for Chinese-speaking learners of Japanese. In *Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation*, pages 296–302. The National University (Phillippines).

Jun Liu, Hiroyuki Shindo, and Yuji Matsumoto. 2018b. Sentence Suggestion of Japanese Functional Expressions for Chinese-speaking Learners. In *Proceedings of ACL 2018, System Demonstrations*, pages 56–61, Melbourne, Australia. Association for Computational Linguistics.

Qianchu Liu, Fangyu Liu, Nigel Collier, Anna Korhonen, and Ivan Vulić. 2021a. MirrorWiC: On Eliciting Word-in-Context Representations from Pretrained Language Models. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 562–574, Online. Association for Computational Linguistics.

Qianchu Liu, Edoardo Maria Ponti, Diana McCarthy, Ivan Vulić, and Anna Korhonen. 2021b. AM2iCo: Evaluating word meaning in context across low-resource languages with adversarial examples. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7151–7162, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alessandro Moschitti. 2006. Making tree kernels practical for natural language learning. In *11th conference of the European Chapter of the Association for Computational Linguistics*, pages 113–120.

Nakamachi, Toshinori, Nishiuchi, Masayu, and Oku. 2022. Estimation of japanese text difficulty based on the japanese language proficiency test. *Online*.

Manabu Okumura, Kiyoaki Shirai, Kanako Komiya, and Hikaru Yokono. 2011. On SemEval-2010 Japanese WSD Task. *Journal of Natural Language Processing*, 18(3):293–307.

OpenAI. 2023. Gpt-4 technical report.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. 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 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.

Ildikó Pilán, Elena Volodina, and Richard Johansson. 2013a. Automatic selection of suitable sentences for language learning exercises. In *20 Years of EUROCALL: Learning from the Past, Looking to the Future: 2013 EUROCALL Conference Proceedings*.

Ildikó Pilán, Elena Volodina, and Richard Johansson. 2013b. Automatic Selection of Suitable Sentences for Language Learning Exercises. In *20 Years of EUROCALL: Learning from the Past, Looking to the Future*, pages 218–225. Research-publishing.net.

Alessandro Raganato, Tommaso Pasini, Jose Camacho-Collados, and Mohammad Taher Pilehvar. 2020. XL-WiC: A multilingual benchmark for evaluating semantic contextualization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7193–7206, Online. Association for Computational Linguistics.

Kristina Hmeljak Sangawa, T. Erjavec, and Yoshiko Kawamura. 2010. Automated collection of japanese word usage examples from a parallel and a monolingual corpus. In *"Proceedings of eLexicography in the 21st century: New challenges, new applications.*

Osamu Sawada. 2007. Two types of adverbial polarity items in japanese: absolute and relative. In *Proceedings of the 10th Conference of the Pragmatics Society of Japan*.

Burr Settles, Chris Brust, Erin Gustafson, Masato Hagiwara, and Nitin Madnani. 2018. Second language acquisition modeling. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 56–65, New Orleans, Louisiana. Association for Computational Linguistics.

Hiroyuki Shinnou and Minoru Sasaki. 2008. Division of example sentences based on the meaning of a target word using semi-supervised clustering. In *International Conference on Language Resources and Evaluation*.

Mitchell Shortt. 2021. Synthesizing a Japanese-language functional expression learning system with Chinese-speaking learners' cultural interests and backgrounds. *Educational Technology Research and Development*, 69(1):319–322.

Patrick E. Shrout and Joseph L. Fleiss. 1979. Intraclass correlations: Uses in assessing rater reliability. *Psychological bulletin*, 86(2):420–428.

Arseny Tolmachev and Sadao Kurohashi. 2017. Automatic extraction of high-quality example sentences for word learning using a determinantal point process. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 133–142, Copenhagen, Denmark. Association for Computational Linguistics.

Arseny Tolmachev, Sadao Kurohashi, and Daisuke Kawahara. 2022. Automatic japanese example extraction for flashcard-based foreign language learning. *Journal of Information Processing*, 30:315–330.

Haifeng Wang, Jiwei Li, Hua Wu, Eduard Hovy, and Yu Sun. 2023. Pre-trained language models and their applications. *Engineering*, 25:51–65.

Monica Ward. 2017. *ICALL's relevance to CALL*, pages 328–332. Research-publishing.net.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2023a. A survey of controllable text generation using transformer-based pre-trained language models. *ACM Comput. Surv.*, 56(3).

Hengyuan Zhang, Dawei Li, Yanran Li, Chenming Shang, Chufan Shi, and Yong Jiang. 2023b. Assisting language learners: Automated trans-lingual definition generation via contrastive prompt learning. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 260–274, Toronto, Canada. Association for Computational Linguistics.

A Difficulty classifier training and evaluation

Parameter	Value
model	cl-tohoku/bert-base-japanese-v3
tokenizer	model's AutoTokenizer
no. labels	5 *(N1, N2, N3, N4, N5)*
learning rate	2e-5
batch size	8
no. epochs	10
adam β_1	0.9
adam β_2	0.999
adam ϵ	1e-7
weight decay	0.01

Table 5: Summary of training parameters for the difficulty classifier.

Class	Precision	Recall	F1-score	Support
N5	0.88	0.88	0.88	25
N4	0.90	0.89	0.90	53
N3	0.78	0.90	0.84	62
N2	0.71	0.79	0.75	47
N1	0.95	0.77	0.85	73
Macro Avg	0.84	0.84	0.84	260
Weighted Avg	0.85	0.84	0.84	260
Accuracy	0.84			260

Table 6: Metrics on data from the test split from the same data distribution for the difficulty classifier.

Class	Precision	Recall	F1-score	Support
N5	0.62	0.66	0.64	145
N4	0.34	0.36	0.35	143
N3	0.33	0.67	0.45	197
N2	0.26	0.20	0.23	192
N1	0.59	0.08	0.15	202
Macro Avg	0.43	0.39	0.36	879
Weighted Avg	0.43	0.39	0.36	879
Accuracy	0.38			879

Table 7: Metrics on a test set of sentences from the official JLPT exams for the difficulty classifier.

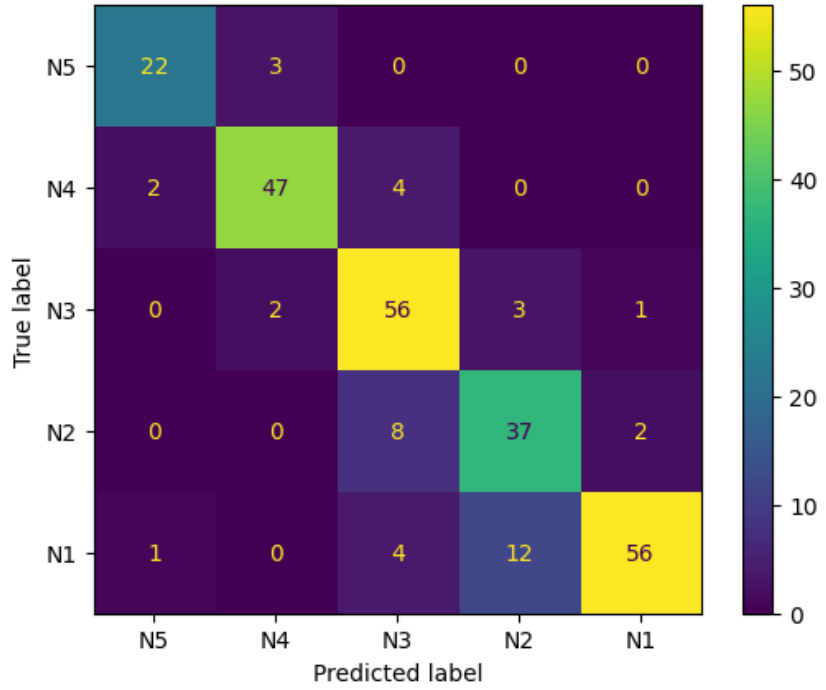

Figure 3: Confusion matrix for the difficulty classifier, on sentences obtained in the same way as the training data (i.e. distant supervision labeling from language websites).

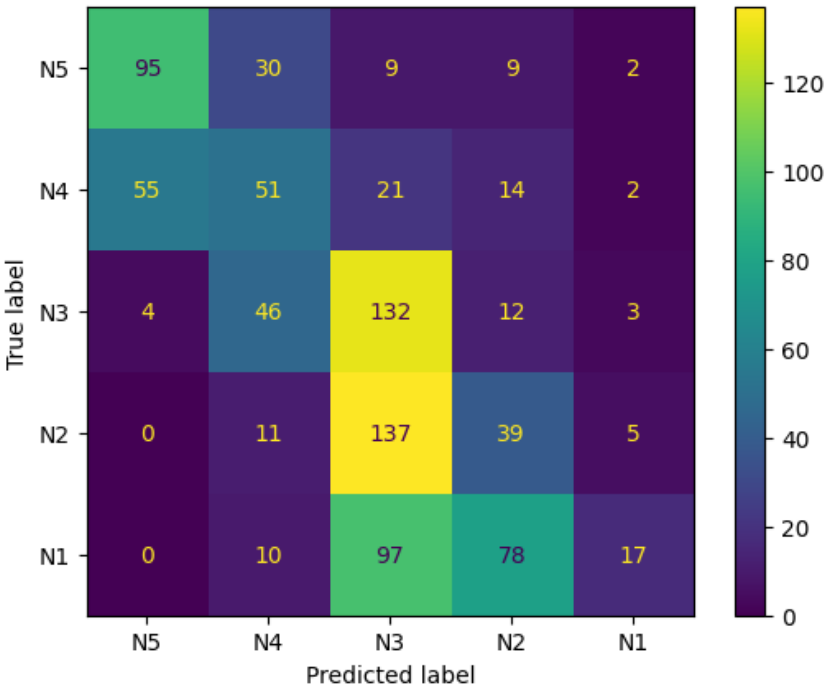

Figure 4: Confusion matrix for the difficulty classifier, on sentences obtained from a different source (i.e. past exams from the official JLPT website).

B Human evaluation form - Example of evaluation block

Context sentence	Target level	Target word	Block ID								
また、東西お互いに相手を非難するプロパガンダ放送を流し合っていた。	N1	相手	1								
System 1				System 2				System 3			
Suggested Sentences	Difficulty rating	Sense rating	Reject	Suggested Sentences	Difficulty rating	Sense rating	Reject	Suggested Sentences	Difficulty rating	Sense rating	Reject
東京と大阪はライバル同士であるため、それぞれの地域では互いに相手を非難するプロパガンダ放送を流し合っていた。	N2	Similar	☐	外交経験が無い素人2人組の外交は半年が空費され、相手から一方的に条件を呑まされる寸前になり失敗に終わった。	N1	Similar	☐	彼は相手の発言に敏感に反応し、即座に的を得た反論を返した。	N3	Similar	☐
これまでずっと、両国間では互いに相手を非難するプロパガンダ放送を流し合ってきた。	N2	Similar	☐	ドイツの戦略爆撃機とイギリス、アメリカの戦略爆撃機の合計の多い国家が少ない国家（ドイツ又はイギリス）を攻撃することになり、相手国家の工業力を低下させる。	N1	Similar	☐	相手の心情を察しつつ、適切なアドバイスを提供することが大切です。	N3	Similar	☐
彼と彼女の関係がうまくいかないときは、私たちは常に相手を責める。	N3	Similar	☑	これはその言葉を発した側が、その発言を持って相手を貶めようとしているためである。	N2	Similar	☐	ディベートでは相手の弱点を見つけ、巧妙に攻撃することが求められます。	N3	Similar	☐
だから、彼らは互いに相手を非難するプロパガンダ放送を流し合っていた。	N2	Similar	☐	カトコフの主張は、一般的に穏健なものではあったが、ひとたびや筆を執るや否や、痛烈に相手を批判せずにはいられなかった。	N1	Similar	☐	相手の意図を読み取りながら、戦術的な効果的な問いかけをする必要があります。	N3	Similar	☐
私たちは互いに相手を尊重し合わなければならない。	N4	Similar	☐	また、両派ともに相手の絶滅を主張し、小型の出刃包丁やハンマーなどを使用した襲撃を続けたため逮捕者や離脱者が多数出て、組織の維持すら危うくなった。	N1	Similar	☐	相手の動きを見極めて、適切なタイミングで反応することが勝利の鍵です。	N3	Similar	☐
Syntactic Diversity				Syntactic Diversity				Syntactic Diversity			
Medium				High				High			
System ranking				System ranking				System ranking			
3rd				1st				2nd			

C LLM baselines prompts

We share the prompts, obtained with manual testing and trial and error. We found that the models responded in a satisfactory way also to prompts where the request was formulated in plain English, as well as in Japanese.

For LLM-jp, this was the prompt used to obtain the final outputs:

> write k target level example sentences in japanese, that must contain the word "target word" used in a similar sense as "context sentence". following are k diverse sentences that must use "target word":

For GPT-3.5, we used the same prompt as the other LLM, and only appended the following instruction to reduce verbosity.

> Provide sentences in Japanese in a numbered list, without any translation or romaji.

D GPT-4 evaluation prompt

We present the prompt given to GPT-4 when rating evaluation blocks with the baselines outputs:

> This evaluation aims to rate and compare three systems in providing good example sentences for learners of Japanese at different proficiency levels. An annotation block consists of proposed sentences by 3 systems for a target word, a context sentence and a target difficulty level. The lists of sentences are supposed to help language learners to see diverse examples of a target word in context.
>
> Difficulty: Rate the difficulty of each sentence according to the JLPT (Japanese Language Proficiency Test) scale, where N1 is the most difficult and N5 is the easiest. Indicate which level a sentence belongs to (one of N1, N2, N3, N4, N5). It is possible that for the target level, the system proposes a sentence that is of a different level (higher or lower). Below is a summary of the proficiency levels.[12]

Level	Description
N1	Complex and abstract Japanese across various contexts.
N2	Everyday Japanese in varied situations, with clear materials on different topics.
N3	Japanese in common everyday situations.
N4	Basic Japanese understanding, including familiar topics, basic vocabulary, and kanji.
N5	Fundamental Japanese, including hiragana, katakana, and basic kanji.

> Sense Similarity: Indicate if the target word in each sentence maintains a close sense as in the original context. Possible values: "similar", "not similar". Think broadly and intuitively, rather than strictly by dictionary definitions.
>
> Reject: For each sentence, indicate "Reject" if you think the sentence is not good or useful (for example because it does not reflect natural use).
>
> Sentence diversity: For each system output list, rate the sentences diversity, focusing on the amount of different uses of syntax and structure. Possible values: "Low", "Medium", "High".
>
> System ranking: Rank the systems' outputs from best to worst, considering the overall usefulness of the example sentences for that word, for a language learner of that proficiency level.
>
> Comment: Leave a short comment.

[12] Taken from https://www.jlpt.jp/e/about/levelsummary.html. The description are put into a table for readability.

E Additional rating statistics

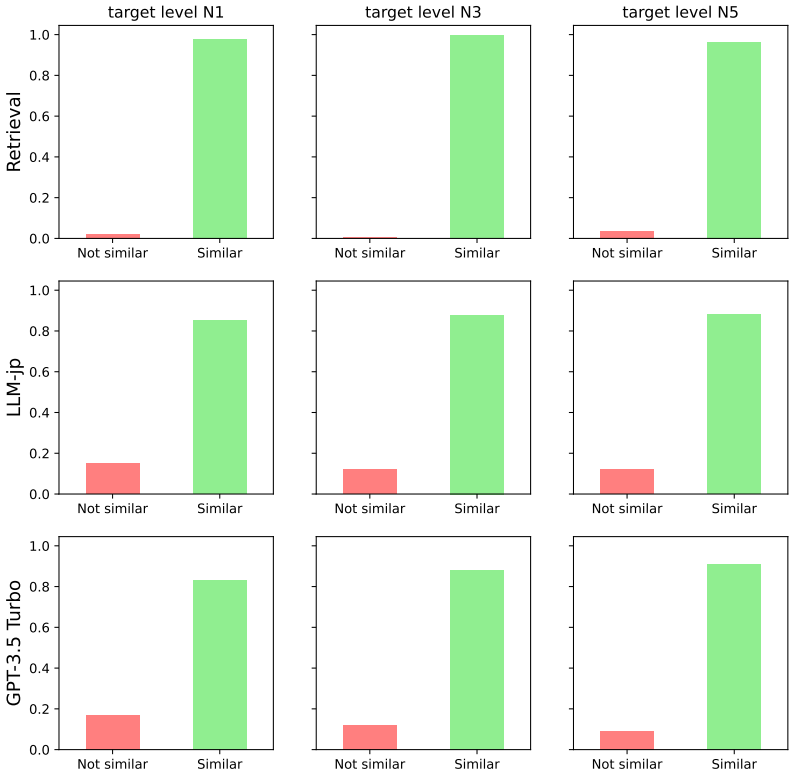

Figure 5: Ratings on sense similarity of proposed sentences.

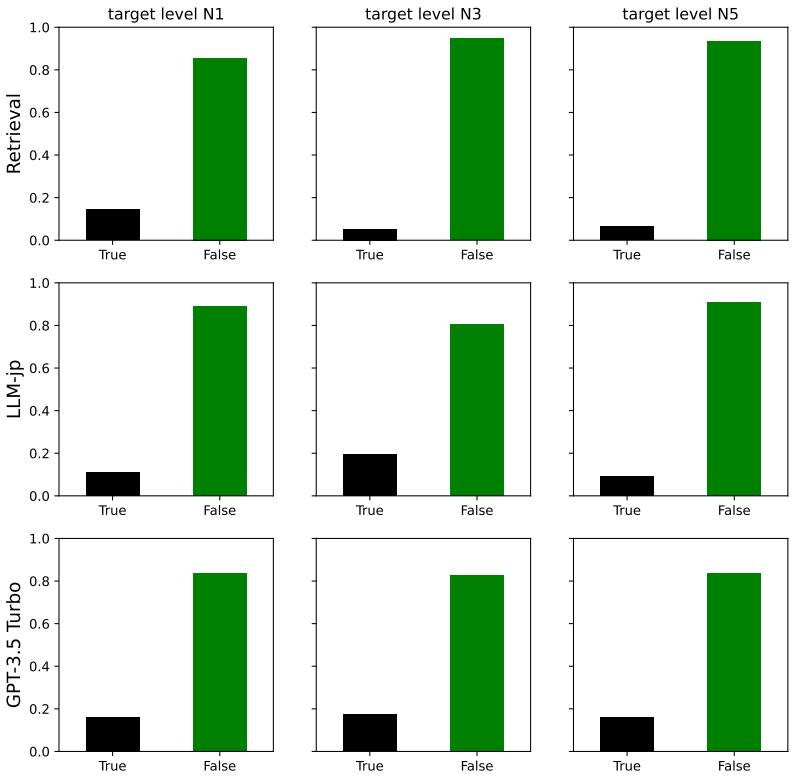

Figure 6: Proportion of rejected proposed sentences.

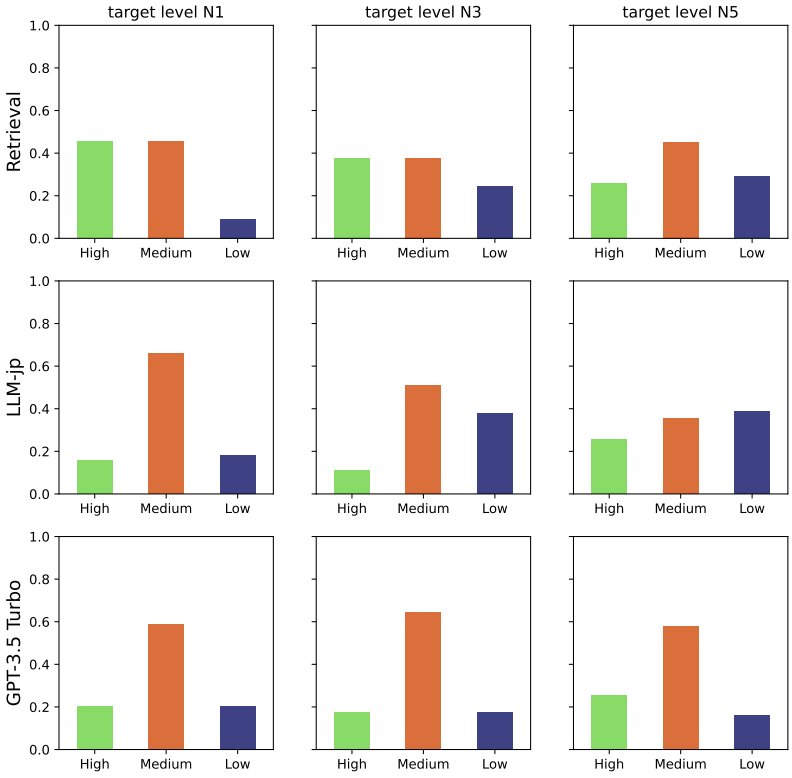

Figure 7: Ratings on syntax diversity of proposed sentences.

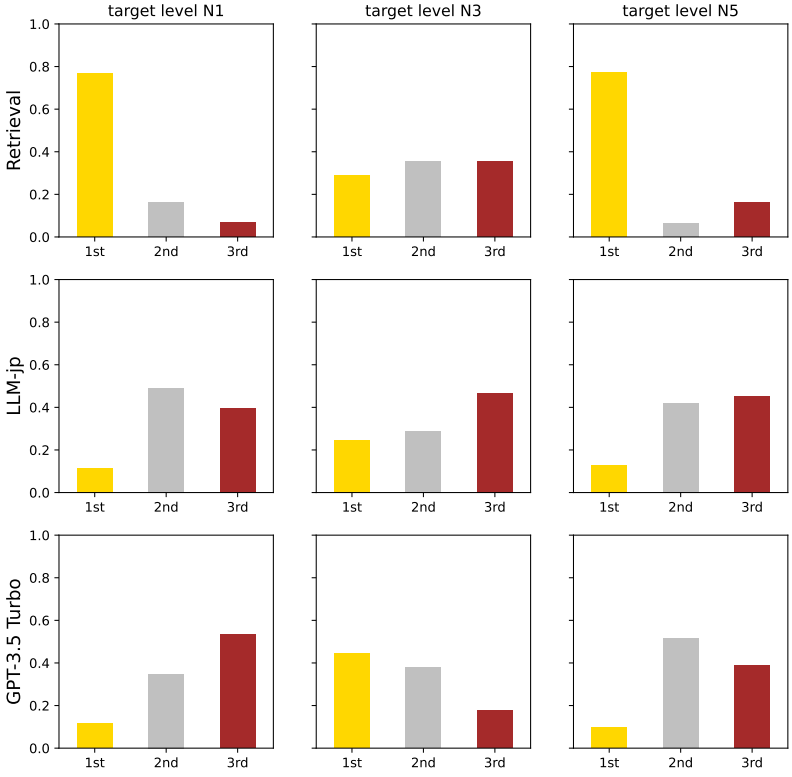

Figure 8: Rankings (first, second, third place) for each system.