Out-of-the-Box Graded Vocabulary Lists with Generative Language Models: Fact or Fiction?

David Alfter
Gothenburg Research Infrastructure in Digital Humanities (GRIDH)
University of Gothenburg, Sweden
first.last@gu.se

Abstract

In this paper, we explore the zero-shot classification potential of generative language models for the task of grading vocabulary and generating graded vocabulary lists. We expand upon prior research by testing five different language model families on five different languages. Our results indicate that generative models can grade vocabulary across different languages with moderate but stable success, but producing vocabulary in a language other than English seems problematic and often leads to the generation of non-words, or words in a language other than the target language.

1 Introduction

Vocabulary lists have long been a cornerstone in language learning, offering learners a structured approach to building their vocabulary and improving reading comprehension (Laufer, 2006; Webb and Nation, 2017; Miralpeix and Muñoz, 2018). Resources like the Academic Word List (AWL; Coxhead 1998) and the New General Service List (NGSL; Brezina and Gablasova 2015) have proven useful for both learners and teachers.

Graded vocabulary lists are a subset of vocabulary lists that include a *grade* for each vocabulary item, indicating its *difficulty* level for learners. This information empowers learners to understand words at their current level, build their vocabulary progressively, and improve their reading skills. For teachers and curriculum developers, graded lists are essential tools for lesson planning and textbook creation, ensuring learners encounter vocabulary appropriate for their proficiency level (Kilgarriff et al., 2014). The importance of graded vocabulary lists is especially clear in the second language learning (L2) context. They are used in language assessment tests (Coxhead, 2011), as vocabulary learning strategies (LaBontee, 2019), in automated essay grading systems (Pilán et al., 2016; Wilkens et al., 2022), in text simplification systems (Tack et al., 2016; Yancey and Lepage, 2018), for automatic exercise generation (Alfter et al., 2019; Alfter and Graën, 2019), to search for appropriate reading materials (Lee and Yeung, 2018; Ehara et al., 2018), or in intelligent tutoring systems (Avdiu et al., 2019).

While graded vocabulary lists have undeniable value, they also come with some limitations. Static vocabulary lists can become outdated as language evolves, and they cannot dynamically adjust to individual learner needs. Furthermore, compiling graded vocabulary lists often requires access to specific textbooks or learning materials, which may not always be readily available or affordable.

The emergence of Generative Language Models (GLMs) presents a potential paradigm shift (Creely, 2024; Godwin-Jones, 2024). These models have demonstrated impressive capabilities in tasks relevant to the L2 context. For example, GLMs can generate difficulty-adapted definitions for words (Kong et al., 2022; Yuan et al., 2022), which helps learners with unfamiliar words. ; simplify complex texts and tailor the difficulty to the learners' needs (Baez and Saggion, 2023); assess essays and provide feedback (Bannò et al., 2024); and perhaps most importantly, GLMs can generate new texts specifically adapted to different difficulty levels (Bezirhan and von Davier, 2023; Kianian et al., 2024; Zualkernan and Shapsough, 2024).

While GLMs hold immense promise, approaching or surpassing human-level performances in some areas (for example in cloze tasks; Rego Lopes et al. 2024), they are not without their drawbacks. Some studies show that current models do not yet outperform task-specific models (Kocoń et al., 2023), that they struggle with vocabulary in an L2 setting (Farr, 2024; Żerkowska, 2024) and lexical complexity prediction (Kelious et al., 2024). Additionally, achieving optimal results with GLMs often requires significant computational resources,

David Alfter. Out-of-the-Box Graded Vocabulary Lists with Generative Language Models: Fact or Fiction?.
Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2024). Linköping Electronic Conference Proceedings 211: 1–19.

potentially limiting their accessibility.

However, now that it is possible to train GLMs on consumer GPUs without strategies such as offloading, model parallel, check-pointing (Zhao et al., 2024), the question arises: In the age of GLMs, do we still need graded vocabulary lists? Can end users easily use GLMs for vocabulary grading purposes, and if so, how well do these models perform? In order to shed light on these questions, we formulate and explore the following hypothesis: *GLMs are effective at grading vocabulary.*

Our contributions are:

1. We investigate the utility of generative language models on the task of grading vocabulary for language learners in a zero-shot setting
2. We test five generative language model families on five (European) languages
3. We show that all models show comparable yet underwhelming performance across the five languages

The rest of the paper is structured as follows: Section 2 contextualizes our work and points to the gaps in current research. Section 3 explains the methodology, including data, experimental setup and evaluation criteria. Section 4 presents and discusses the results. Sections 5 and 6 round off the paper with conclusion and future work.

2 Related Work

There are two research strands that are closely connected to this line of research: complex word identification (Paetzold and Specia, 2016) and lexical complexity prediction (North et al., 2023b). Complex word identification is concerned with identifying *complex* words with downstream applications such as lexical text simplification (Shardlow, 2013; Maddela and Xu, 2018). It is a binary task (is a word complex or not), and is not specifically targeting the L2 context.

Lexical complexity prediction emerged from complex word identification and aims at classifying the complexity of words on a *graded* scale (e.g., how complex is a word, on a scale from 1 to 4). Lexical complexity prediction is also mainly used for downstream tasks like text simplification (North et al., 2023a; Shardlow et al., 2024b), and is not specifically targeting the L2 context. However, as demonstrated by the ongoing list of shared tasks on the topic (Paetzold and Specia, 2016; Yimam et al., 2018; Ortiz-Zambranoa and Montejo-Ráezb, 2020; Shardlow et al., 2024a), it is still an active area of research. The latest lexical complexity prediction shared task was a sub-task of the BEA shared task on multilingual text simplification (Shardlow et al., 2024a).

Recent work on complex word identification and lexical complexity prediction found that ChatGPT only sometimes outperforms task-specific models, mostly in cases when the contexts are dissimilar enough to allow for the discovery of a difference; task-specific models tend to perform better at discriminating the complexity of words even with smaller context variations (Kelious et al., 2024). In the recent shared task on multilingual lexical complexity prediction and lexical simplification, the winning team of sub-task 1 (lexical complexity prediction) used GPT4, with an average Pearson correlation of 0.62 (Enomoto et al., 2024).

On the other hand, generative language models and their potential for on-the-fly generation of learning material is increasingly being investigated. However, the focus of these studies is mostly on text passage generation (Attali et al., 2022; Bezirhan and von Davier, 2023; Peng et al., 2023; Boras et al., 2024) and personalization (Leong et al., 2024; Pesovski et al., 2024).

We fill a critical gap in the literature by investigating the potential of GLMs for graded vocabulary lists and by extending the analysis to multiple different models and multiple languages on comparable data.

3 Methodology

In this paper we explore two uses cases for GLMs and graded vocabulary lists. First, we suppose that a researcher/learner/teacher is in possession of an ungraded word list that they might want to grade using GLMs. Second, we suppose that no vocabulary list exists, and the researcher/learner/teacher wants to create a graded vocabulary list from scratch using GLMs. In both cases, we compare the output of the GLMs to existing vocabulary lists, using both qualitative and quantitative evaluations (see Section 3.3 for evaluation criteria).

3.1 Data

As data for this investigation, we use the freely available CEFRLex[1] lists. These lists are derived

[1] https://cental.uclouvain.be/cefrlex

from textbooks aimed at learners of different languages and contain among others for each lemma the frequencies at different textbook levels (see Figure 1) according to the Common European Framework of Reference for Languages (CEFR; Council of Europe 2018). We specifically use EFLLex (Dürlich and François, 2018) for English, ELELex (François and De Cock, 2018) for Spanish, FLELex[2] (François et al., 2014) for French, SVALex for Swedish (François et al., 2016) and NT2Lex (Tack et al., 2018) for Dutch[3].

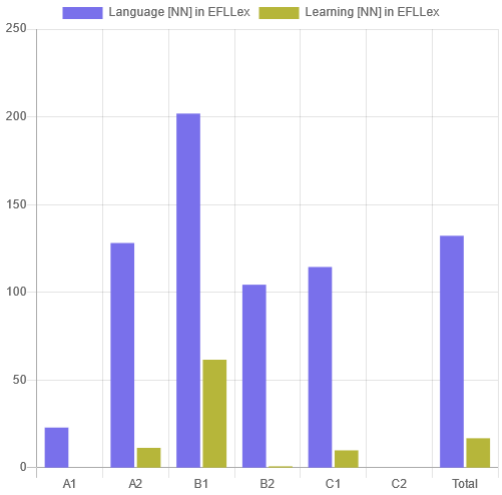

Figure 1: Frequencies across levels for the words 'Language' and 'Learning' in EFLLex

While the Cambridge English Vocabulary Profile (EVP; Capel 2015) or Pearson's Global Scale of English (GSE; Pearson 2017) might potentially be more widely used, they are not available in a machine-readable form, being targeted at human end users. Furthermore, they only cover the English language. However, a study comparing these two resources between themselves and to EFLLex found moderate to high correlations both between EVP and GSE (0.85) and between EVP&GSE and EFLLex (0.70; Graën et al. 2020).

As the word lists contain some artifacts and word fragments (e.g., -hour_day, bly7458/00578, flight_kl0549), we perform some data cleaning. We only retain single words (excluding multi-word expressions), and exclude words that contain non-alphabetical characters such as digits or other symbols. We only retain nouns, verbs, adjectives, and adverbs.

Finally, we map each word to the level at which it is first observed (first-occurrence approach). While simple, this method has been shown to perform on-par with more complex level assignment methods (Gala et al., 2013; Alfter, 2021). We opt for a numerical scale rather than the CEFR scale that the word lists are derived from, mapping A1 to 0, A2 to 1, B1 to 2, B2 to 3, and C1 to 4. We disregard C2, which is only included in the French list, as the difference between C1 and C2 is difficult to assess (Springer, 2012; Sung et al., 2015; Isbell, 2017), and the focus of the study lies less in the discriminatory performance at the highest levels but rather a general ability to grade vocabulary from easiest to hardest.

Table 1 shows an overview over the final word lists used in the experiments.

List	WC	WC2
EFLLex (English)	29667	10295
ELELEx (Spanish)	14290	13291
FLELex (French)	17237	13242
SVALex (Swedish)	15634	13662
NT2Lex (Dutch)	17743	13972

Table 1: Overview over word counts before (WC) and after (WC2) data cleaning

3.2 Experimental setup

We test five popular instruction-tuned model families: Google's Gemma (Gemma Team et al., 2024), MistralAI's Mistral (Jiang et al., 2023), Meta's Llama (Touvron et al., 2023), Microsoft's Phi3 (Abdin et al., 2024), and OpenAI's GPT (OpenAI et al., 2024). Specifically, we use Gemma-1.1-2b-it, Gemma-1.1-7b-it, Mistral-7B-Instruct-v0.3, Llama3-8B-Instruct[4], Phi-3-mini-4K-instruct, and GPT-4o. Gemma, Mistral, Llama, and Phi3 provide small versions of their models (2B to 8B) that do not necessitate massive servers to run, while GPT-4o potentially relies on multiple different models of larger size (cited as exceeding 200B; Ayub 2024) but can be queried programmatically, thus requiring only a paying account and access to the internet.

[2]From the three available versions, for reasons of comparability, we chose the TreeTagger version without automatically assigned CEFR labels.

[3]We do not take into account the sense-disambiguated version of this list, as it mirrors the original list with additional sense labels

[4]Preliminary experiments with Llama2-7b-chat showed a strong underperformance in comparison to the other models, an "unwillingness" to follow instructions, and a tendency to mostly respond with a score of 3. As a result, the model was excluded from further experiments.

All models (except GPT4o) are loaded in 4bit quantized form, and GPT4o is queried through its API. All calculations were performed on a single high-end laptop computer with a 12th Gen Intel®Core i7 2.40Ghz processor, 32GB RAM and an NVIDIA GeForce TRX 3080 Ti Laptop GPU graphic card.

Parameters for the models were taken from their respective Hugginface pages with sample code, mirroring a 'naive' approach to using GLMs by simply copy-pasting their example code and running it. This means that some models use sampling or have a temperature parameter above zero, reducing the reproducibility of this study. All parameters can be found in Appendix B, Table 8.

3.2.1 Generating grades

For the first experiment, we use the word lists as basis and ask the generative language models to grade the vocabulary.

Similar to Enomoto et al. (2024) who prompt GPT4 with a single English prompt for lexical complexity values for different languages, we use a single English prompt for all languages and models, with the first part specified as *system* input if the model supports a *system* role, otherwise prepended to the *user* prompt. The full prompt is:

> You are an experienced teacher of *language* as a second language. You can easily assess the difficulty of words in *language* for learners. You assess words on a scale from 0 (easiest) to 4 (hardest). You only answer with a number.
>
> Assess: *word* (*part-of-speech*)

3.2.2 Generating vocabulary list

For the second experiment, we ask the generative language models to generate word lists from scratch.

Given the generation limit of GLMs and the associated cost, and the more qualitative evaluation of this experiment, we opt to prompt each model for a maximum of 100 words per level, using the following prompt. As the output may include repeated words, we take the set of unique words for each level and compare it to the word lists.

> You are an experienced teacher of *language* as a second language. You can easily tell which words are suitable for learners of *language* at different levels. You assess words on a scale from 0 (easiest) to 4 (hardest).
>
> Generate 100 words for learners of level *level*.

3.3 Evaluation

First, we evaluate the models according to correctness in predicting grades in comparison to the textbook-derived grades assigned by the first-occurrence approach. For this quantitative evaluation, we use Pearson correlation, Jensen-Shannon distance, accuracy, adjacent accuracy (the prediction is considered correct if it deviates from the target level by at most one level), precision, recall, and F1 score.

Second, we evaluate the quality of the generated graded word lists. For this more qualitative evaluation, we consider coverage of generated vocabulary as the overlap with existing word lists and a more in-depth analysis and discussion.

We also investigate whether there is a link between frequency and discrepancy in prediction. A low frequency in the word list means that the level assignment will be less reliable; if we only observe one occurrence of a word, the level of the word will be the level where it was observed, by definition. If GLMs are *consistent* in grading, then we expect them to grade low-frequency words according to their own internal criteria (as opposed to observed frequency). Further, if GLMs are consistent and *correct* in grading vocabulary, then we expect that larger discrepancies are found in words with low frequency, and less discrepancy in high frequency words.

In addition, we explore the impact of the chosen grading scale, investigating whether prompting the models to grade vocabulary on the CEFR scale rather than a numerical scale might improve results. We have opted for a numerical scale because it might be a more generalizable concept for models to work with, rather than the CEFR scale, which the models might have limited knowledge of. For reasons of economy, we only perform this experiment using the best performing model and two word lists: the one it scored worst on, and the one it scored best on.

4 Results and Discussion

In this section, we report the results from the experiments and discuss the results. For space reasons, model names and word list names are abbrevi-

ated, with G2 and G7 standing for Gemma-2B and Gemma-7B respectively, GPT for GPT-4o, L8 for Llama3-8B, M7 for Mistral-7B, and P3 for Phi-3; EN for EFLLex, ES for ELELex, FR for FLELex, SV for SVALex, and NL to NT2Lex.

4.1 Generating Grades

As a first measure of comparison, we compare the predicted label distributions to the original label distribution by normalizing the label counts by the total number of items and applying the Jensen-Shannon distance measure (Lin, 1991). This indicates how well the predictions follow the original label distributions, although it gives no indication of the *accuracy* of predicted labels. Table 2 shows the Jensen-Shannon distance between the original label distribution and the predictions for each model.

	G2	G7	GPT	L8	M7	P3
EN	0.30	0.40	**0.24**	0.43	0.32	0.46
ES	0.31	0.35	0.22	0.33	**0.20**	0.39
FR	0.48	0.51	0.32	0.40	**0.27**	0.40
SV	0.22	0.42	**0.12**	0.30	0.37	0.37
NL	0.47	0.43	0.16	0.32	**0.13**	0.17

Table 2: Jensen-Shannon distance between the original label distribution and the predicted label distributions by model. Results in bold indicate the best result per language.

In order to check for *accuracy*, we calculate accuracy, precision, recall, weighted F1 score, and adjacent accuracy. For reasons of space, we only report F1 scores in the main body of the paper. The full table including accuracy, adjacent accuracy, precision, and recall, can be found in Appendix A, Table 7. Table 3 shows the weighted F1 scores for each model and word list.

	G2	G7	GPT	L8	M7	P3
EN	0.17	0.18	**0.29**	0.16	0.24	0.15
ES	0.15	0.19	0.24	0.20	**0.28**	0.19
FR	0.15	0.12	0.21	0.19	**0.28**	0.22
SV	0.26	0.30	**0.33**	0.25	0.18	0.20
NL	0.18	0.19	0.35	0.35	0.36	**0.38**

Table 3: Results in terms of Weighted F1 score. Results in bold indicate the best result per language.

For comparability to lexical complexity prediction, we also calculate Spearman's ρ.[5] Table 4 shows the results.

	G2	G7	GPT	L8	M7	P3
EN	0.03	0.29	**0.48**	0.36	0.40	0.38
ES	-0.03	0.22	**0.42**	0.29	0.33	0.26
FR	0.03	0.29	**0.46**	0.33	0.39	0.37
SV	0.07	0.22	**0.39**	0.25	0.25	0.29
NL	0.07	0.24	**0.38**	0.26	0.27	0.32

Table 4: Results in terms of Spearman's ρ. Results in bold indicate the best result per language.

Both tables 2 and 3 show a similar trend, with GPT-4o performing best on English and Swedish, Mistral performing best on Spanish and French, and Mistral performing best on Dutch in terms of predicted label distribution but outperformed by Phi-3 in terms of weighted F1 score. Table 4 shows that GPT-4o correlates most with the reference data in all cases, followed by Mistral-7B and Phi3.

Interestingly, although most models are exclusively meant for use with the English language, all models show a rather good cross-linguistic capacity. Further, none of the models performed particularly well in English, or remarkably better on English in comparison to the other languages.

Given possible fluctuations, it seems that both Mistral-7B and GPT-4o are performing similarly well on this task. Given that GPT-4o requires a paying subscription, Mistral-7B seems to be a viable free alternative. We can also observe that Mistral-7B performs quite well across languages, except for Swedish, where the Gemma models show surprisingly good performance, coming second (G7) and third (G2) after GPT-4o. We can also observe that all models except the Gemma family performed best on Dutch. Finally, we may see a language bias: Mistral-7B performed best on Romance languages, while GPT-4o performed best on Germanic languages, potentially reflecting a bias in training data.

4.2 Frequency and Discrepancy

For this experiment, we order each list by total frequency as given in CEFLex and calculate the absolute difference in predicted level and assigned

[5]The lexical complexity prediction tasks indicate both Spearman's ρ and Pearson's correlation coefficient, since the numerical labels can be expressed as continuous numbers. However, we do not assume a normal distribution of the data, which is a prerequisite for Pearson's correlation coefficient.

level. We then calculate the average discrepancy for the first and last x entries, varying x from 10 to 100 in steps of 10. Figure 2 shows the discrepancy for the different values of x.

As can be seen from the figure, Gemma-2B shows an opposite trend of what would be expected with higher discrepencies for high frequency words, and lower discrepancies for low frequency words. Gemma-7B shows a mixed picture, with the expected trend at $x = 10, 20, 30$, but an opposite trend from $x = 40$. GPT-4o, Llama3-8B, Mistral-7B, and Phi3-4K display a higher discrepancy for the lowest frequency words and a lower discrepancy for the most frequent words across all languages, following the expected pattern and confirming that GLMs may be useful for grading vocabulary items for which the total observed frequency is too low.

4.3 Impact of Grading Scale

As noted previously, we only investigate the impact of the grading scale using the best model and the word lists it performed best and worst on. Based on Table 2, we select Mistral-7B as model and Swedish and Dutch as word lists. For the two word lists, we proceed as described in Section 3.2.1, but we modify the prompt as follows:

> You are an experienced teacher of *language* as a second language. You can easily assess the difficulty of words in *language* for learners. You assess words on **the CEFR scale ranging from A1 (easiest) via A2, B1, B2, to C1 (hardest). You only answer with a CEFR label.**
>
> Assess: *word* (*part-of-speech*)

	Numerical scale	CEFR scale
SV	0.18	0.12
NL	0.36	0.20

Table 5: F1 scores (weighted) for numerical scale and CEFR scale

Table 5 shows the comparison between using a numerical scale versus using the CEFR scale. We can note a marked decrease in performance for both word lists, hinting at the possibility that the language model may not have come into contact with the CEFR in sufficient amounts to be able to accurately apply it. We also notice a tendency towards predicting A1, which may be due to the problem of *primacy*, a tendency for the model to pick the first alternative from a list of alternatives, previously shown to exist in ChatGPT (Wang et al., 2023).

4.4 Generating Vocabulary Lists

In this section we present the results of the vocabulary generation task. During result examination, we noticed that Gemma-7B consistently output numbered lists that only list items 1-10 and 90-100, with ellipsis of the rest. We therefore opted to leave out the results for Gemma-7B in this section.

Table 6 shows an aggregated version over all languages and all levels for vocabulary generation. The table shows that we requested 2500 words from each GLM, with 100 words distributed over five levels for five languages ($100 * 5 * 5$). We can see that only GPT-4o generated the exact number of requested words, Llama-3 generated almost the requested number of words, Gemma did not provide even half of the requested words, while Mistral-7B and Phi-3 overgenerated. However, the generated vocabulary lists contain duplicates. Based on the unique count of words, we can see that GPT-4o was closest to the target, followed by Llama-3 (who overgenerated).

When looking at the number of items generated at the requested level, we can again see that GPT-4o performed best, followed by Mistral-7B. However, Mistral-7B also shows the highest out-of-vocabulary rate, meaning that it generates words that are not present in the reference word list. In terms of overall coverage, we can see that GPT-4o performs best, followed by Mistral-7B and Llama-3-8B.

A detailed investigation of results reveals that Mistral-7B and Llama3-8B tend to group words by categories (numbers, days of the week, months of the year, greetings, travel, family, weather,...). Gemma often disregards the requested level and generates a list spanning all levels, grouped by level (easy, moderate, challenging, complex); this behavior is sometimes also observed for Mistral (French and Spanish). Phi3 does generate a list of at least 100 items, but starts repeating the same word after 20-30 words.

In the following, we examine each model language by language and investigate the causes for a low overlap by looking at words that the model generated that were not found in the reference list,

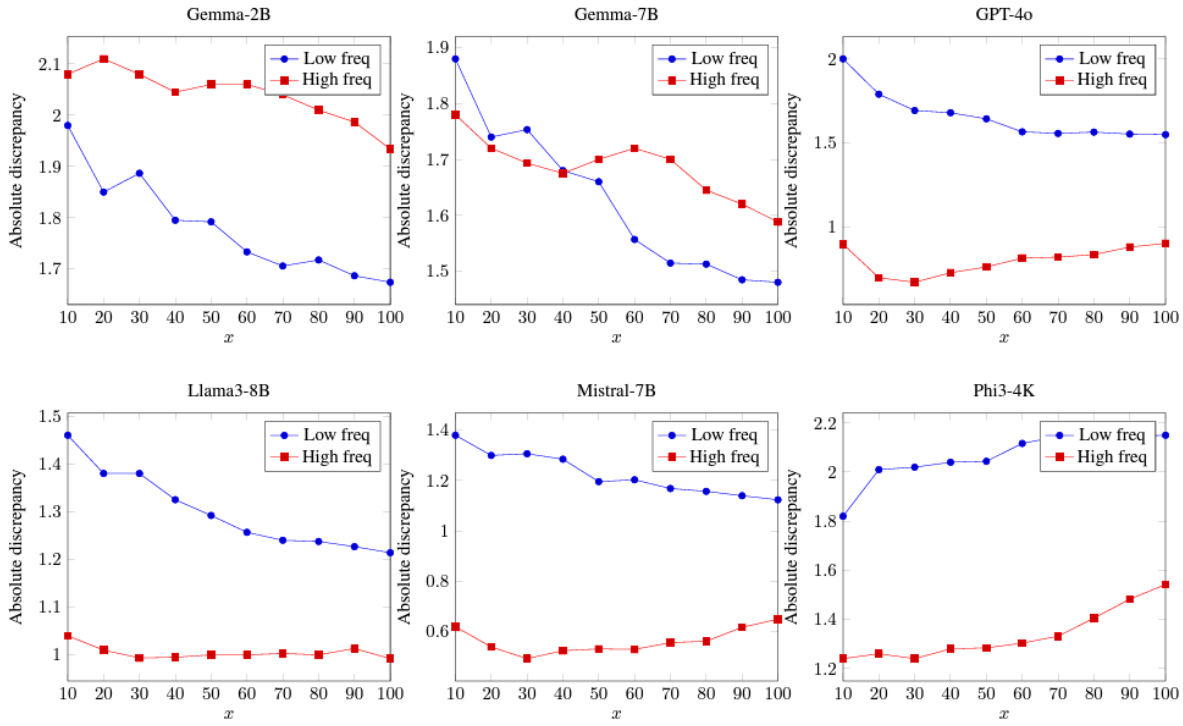

Figure 2: Discrepancies over different values of x. Each graph shows the average absolute discrepancy over all languages for the most frequent (High freq) and least frequent (Low freq) words when taking into account the x first and last words of the frequency-ranked list.

	R	G	U	L	D	OOV	OOVR (%)	LC (%)	OC (%)
Gemma-2B	2500	1204	935	156	542	237	25.35	6.24	27.92
Llama3-8B	2500	2433	2181	283	1148	750	34.39	11.32	57.24
GPT-4o	2500	2500	2460	487	1448	525	21.34	19.48	77.40
Mistral-7B	2500	2675	2574	355	1086	1133	44.02	14.20	57.64
Phi3-4K	2500	3073	1285	202	579	504	39.22	8.08	31.24

Table 6: Coverage of generated vocabulary lists aggregated over all languages and levels, with the requested number of words (100 per level per language; R), the number of generated words (G), the number of unique generated words (U), the number of words generated that correspond to the desired level (L), the number of generated words that are in the word list but at a different level (D), the number of generated words that are not in the reference word list (Out-of-vocabulary; OOV), the out-of-vocabulary rate (out of all generated unique words, how many are not in the reference word list; OOVR), the level coverage (out of all generated words, how many are in the reference word list at the given level; LC), and the overall coverage (out of all generated words, how many are in the reference word list; OC)

then draw an overarching picture.

4.4.1 English

Gemma-2B At level 0, the generated words include conjunctions (or) and prepositions (from) that were excluded from the reference word list due to their part-of-speech. At level 1, the model generates pronouns (I), interjections (hello) and multi-word expressions (family member, thank you) that were also excluded from the reference word list. At level 2, the model generates words that, according to the authors of the paper, are of arguably higher complexity than level 2 (e.g., accountability, resilience, transformative). At level 3, the model generates plausible candidates. At level 4, the model generates words that, again, are arguably above level 4, such as: superimpose, reticent, parsimonious, abrogate, concomitant, magnanimous, prevaricate, obsequious, iconoclast.

GPT-4o At level 0, the model includes personal pronouns (you, us) and numbers (zero) that were excluded based on part-of-speech. However, the model also generates some words that are suitable but missing from the reference list (lion, ant). At level 1, the model generates interjections (hello). At level 2, the model generates months of the year and prepositions (between, during). At level 3, the model generates plausible candidates. At level 4, the model generates words that are plausible but includes words of arguably higher complexity, such as: pernicious, surreptitious, vicissitude, obstreperous, prevaricate

Llama3-8B At level 0, the model generates numbers (four), multi-word expressions (thank you) and plural forms (socks). At level 1, the model generates words that would potentially be more appropriate at level 0 (lion, rectangle, triangle). At level 2, the model generates plural forms (nuts, pillows), easier words (lemon, omelette) but also plausible candidates. At level 3, the model generates words of a much higher level (inscrutable, garrulous, obfuscate, sagacious, jaded, callipygian). At level 4, the model generates even more complex words (abstruseness, papaphobia, mumpsimus, insouciant, tintinnabulation, perspicacious, zephyrine, gymnosophy).

Mistral-7B At level 0, the model generates numbers (zero), prepositions (among, between, from) and question particles (who, where, when, why, what) that were excluded from the reference list based on part-of-speech. At level 1, the model generates interjections (hello), multi-word expressions (last week, thank you, next week, wake up), and conjugated verb forms (does, hasn't). At level 2, the model generates numbers (four, six) and multi-word expressions (I am fine, what time is it?, I do, you're welcome). At level 3, the model generates words that are of much higher complexity (jocund, aphorism, temerity, blithe, capricious, komphetamology). At level 4, the model also generates words of much higher complexity (obstreperous, sanctimonious, capacious, lachrymose).

Phi3-4K At level 0, the model generates numbers (seven, nine, four, six), multi-word expressions (a lot, thank you), but also some plausible missing words (giraffe, kangaroo, lion). At level 1, the model generates plural forms (shoes, socks, pants) and multi-word expressions (thank you, living room). At level 2, the model generates plural forms (shoes, socks, pants) and multi-word expressions (thank you). At level 3, the model generates plausible words but also words with arguably higher complexity (xylography, opulence). At level 4, the model generates plausible words but arguably of higher complexity (obfuscation, zephyr, rambuctious, nebulous, taciturn, dichotomy, ephemeral, ineffable, effulgent, limerence).

4.4.2 Spanish

Gemma-2B At level 0, the model generates good candidates that simply are not in the reference word list (día 'day', gato 'cat', perro 'dog', casa 'home'). MWE: gracias de nuevo, por favor too high level: inspirador, felizmente number: uno At level 1, the model generates numbers (uno 'one'), multi-word expressions (gracias de nuevo 'thanks again', por favor 'please'), but also words of a higher complexity (inspirador 'inspiring/inspirer', felizmente 'happily'). At level 2, the model generates plausible candidates. At level 4, the model also generates plausible candidates, although one word seems to be misspelled (*objetovo, probably objetivo '(an) objective').

GPT-4o At level 0, the model generates numbers (nueve 'nine', tres 'three'), feminine forms (hermana 'sister') and multi-word expressions (por favor 'please'). At level 1, the model generates interjections (hola 'hello', gracias 'thanks'), as well as *hermana* and *por favor* from the previous level. At levels 2 and 3, the model generates plausible words. At level 4, the model generates plausible words but also words with a higher complexity (caliginoso 'caliginous', inasible 'ungraspable', imperecedero 'imperishable', impertérrito 'undaunted').

Llama3-8B At level 0, the model generates some good words (computadora 'computer', telefono 'telephone'), but also some plural forms (animales 'animals', recursos 'resources', familias 'families') and words of higher complexity (pormenor 'detail'). At level 1, the model generates plural forms (llaves 'keys', piernas 'legs', brazos 'arms', manos 'hands'), *hermana*, interjections (hola 'hello'), months of the year, days of the week and numbers. At level 2, the model still proposes *computadora*, *hermana*, and *esposa* 'wife'. At level 3, the model generates multi-word expressions (lo siento 'I'm sorry', me gustaria 'I would like to have', hasta luego 'see you soon'). At level 4, the model generates multi-word expressions and phrases, albeit with only few base constructions, such as *desarrollar* 'develop ' (estrategias 'strategies', habilidades 'habits', ...), and *enfermadad de* 'disease of' (alzheimer 'Alzheimer', cuidados intensivos 'intensive care', ...).

Mistral-7B At level 0, the model generates a lot of words with articles (el choche 'the car', la nariz 'the nose', *el nariz, el pantalón 'the pants', el sombrero 'the hat', el diente 'the tooth', la boca 'the mouth') and multi-word expressions (¿donde está el parque? 'where is the park?', ¿como se dice en español? 'how do you say this in Spanish?', me gusta 'I like'. At level 1, the model generates numbers, interjections (hola 'hello'), conjunctions (con 'with'), conjugated verbs (ríe 'laughs', llora 'cries'), and multi-word expressions (lo siento 'I'm sorry', no me gusta 'I don't like'). At level 2, the model generates numbers, plural forms (amigos 'friends', aguas 'waters') and multi-word expressions (buenas tardes 'good evening', buenas noches 'good night'). At level 3, the model generates plausible words, but also plural forms (familiares 'familiar-ADJ-PL', misterioses 'mysterious-ADJ-PL', hombres 'men', tiempos 'times', equipos 'teams', ventajas 'advantages') and conjugated verb forms (mantiene 'maintains', empieza 'begins', hablaste 'you spoke', cómprame 'buy me!'). At level 4, the model generates mostly plausible words but also French words (flâner 'stroll around') and words with higher complexity (zozobrar 'capsize', cenotafio 'cenotaph', panoptico 'panoptic', acriminarse 'incriminate onself').

Phi3-4K At level 0, the model generates interjections (hola 'hello', gracias 'thanks'), multi-word expressions (a veces 'sometimes', por favor 'please') and plural forms (olas 'waves'). At level 1, the model generates multi-word expressions (manazana roja 'red apple', manzana amarilla 'yellow apple', manzana verde 'green apple'), interjections (hola, gracias) and multi-word expressions (por favor). At level 2, the model generates interjections (hola, gracias). At level 3, the model generates personal pronouns (nosotros 'us'), plural forms (olas 'waves', mesas 'tables', pájaros 'birds'), conjugated verb forms (llegaron 'they arrived', llegaste 'you arrived', llego '(I) arrive'). At level 4, the model generates multi-word expressions (nave espacial 'spacecraft', cambio climático 'climate change', historia antigua 'old history', jardín botánico 'botanical garden', naturaleza muerta 'still life').

4.4.3 French

Gemma-2B At level 0, the model generates feminine forms (grande 'tall-FEM', petite 'small-FEM'), interjections (oui 'yes'), plural forms (amis 'friends') and multi-word expressions (merci beaucoup 'thank you very much'). At level 1, the model generates *grande* as on the previous level. At level 2, the model generates plausible words. At level 3, the model generates *oui* as on level 0. At level 4, the model generates feminine forms (ambigüe 'ambiguous-FEM') and apparently English words (incoherence, discreet).

GPT-4o At level 0, the model generates interjections (excusez-moi 'excuse me', oui 'yes'), multi-word expressions (s'il vous plaît 'please', au revoir 'goodbye'), feminine nouns (amie 'friend-FEM') but also slightly misspelled words (velo instead of vélo 'bicycle'). At level 1, the model generates numbers, plural forms (amis 'friends'), multi-word expressions (l'année *derniere 'last year'), but also words of lesser complexity (mois 'month'). At level 2, there are no generated words not present in the reference word list. At level 3, the model generates multi-word expressions/reflexive verbs (se faufiler 'sneak'). At level 4, the model generates plausible words, but possibly of too high complexity (prestidigitation 'sleigh of hand', pugnacité 'pugnacity', malversation 'embezzlement', acquiescer 'acquiesce').

Llama3-8B At level 0, the model generates mostly multi-word expressions and phrases or phrasal fragments (je suis impatient 'I am impatient', je voudrais 'I would like', c'est faux 'that's wrong'), but also some questionable phrases such as *ça est irraisonable*, which should be *c'est irraisonable*. At level 1, the model again mostly generates phrases, and again some questionable phrases such as *je suis*

frère/femme ‘I am brother/woman’. At level 2, the model generates plausible words and multi-word expressions (réservation de taxi ‘taxi reservation’, transport en commun ‘public transport’). At level 3, the model generates plausible words and plural forms, although these are generally encountered in the plural (chaussures ‘shoes’, souliers ‘shoes’, épices ‘spices’). At level 4, the model generates some questionable English words of high complexity as mantic, catharsis, and kibosh.

Mistral-7B At level 0, the model generates personal pronouns (eux ‘them’, elle ‘she’), multi-word expressions (pommes frites ‘French fries’), but also some clearly non-French words (beef, chicken, vino). At level 1, the model generates conjugated verb forms (parlait ‘(s/he) spoke’), plural forms (doigts ‘fingers’), multi-word expressions (au revoir ‘goodbye’), and some questionable or wrong forms such as *s’lever* (possible in slang but generally *se lever*), ecouter (écouter), cafe (café). At level 2, the model generates plural forms (chiens ‘dogs’) and some questionable words such as **prenon*, coche ‘car-SPANISH’, milk, banana, egg, water. At level 3, the model generates feminine forms (délicieuse ‘delicious-FEM’) and English words (negociate). At level 4, the model generates multi-word expressions (une fois de plus ‘once more’, penser qu’il est possible ‘think that it is possible’, selon une étude ‘according to a study’), feminine forms (contemporaine ‘contemporary-FEM’), English words (idiosyncrasy), and questionable so-called “multi-word expressions” (trouver des choux de bruxelles sous les pierres ‘finding brussels sprouts under stones’, donner sa bague à quiconque veut l’attraper ‘giving your ring to anyone who wants to grab it’, s’asseoir sur *un *chais de poule ‘sitting on a chicken chair(?)’).

Phi3-4K At level 0, the model generates multi-word expressions (très bien ‘very good’, je n’ai pas ‘I don’t have’, je suis ‘I am’, je ne comprends pas ‘I don’t understand’, pas mal ‘not bad’). At level 1, the model generates male/female alternatives (apprenti(e) ‘apprentice’, professeur(e) ‘professor’, *enfant(e)[6]), multi-word expressions (je suis ‘I am’, très bien, merci ‘very good, thank you’, je m’appelle ‘my name is’, s’il vous plaît ‘please’). At level 2, the model generates multi-word expressions (j’ai besoin de ‘I need’, un peu ‘a bit’, je voudrais ‘I would like’). At level 3, the model generates plural forms (conséquences ‘consequences’, héros ‘heroes’), multi-word expressions (justice sociale ‘social justice’, liberté individuelle ‘individual liberty’), and English words (warrant). At level 4, the model generates plausible words.

[6] *Enfant* as a noun can take both the male and female article. *Enfante* exists as a conjugated form of *enfanter* ‘to give birth/bear fruit/bear a child’

4.4.4 Swedish

Gemma-2B At level 0, the model generates superlative adjectives (bästa ‘best’, högsta ‘highest’), conjugated verb forms (kom ‘come-IMP/came’), alternatives separated by slash (ja/nej ‘yes/no’). At level 1, the model generates more alternatives separated by slash (goddag/godnatt ‘good day/night’, skapar/tar ‘create/take’, jag/du/han/hon ‘I/you/he/she’). At level 2, the model generates personal pronouns (du ‘you’, vi ‘we’), words of higher complexity (semantik ‘semantics’, multipel ‘multiple’, konnotation ‘connotation’) and conjunctions (som ‘as’). At level 3, the model generates non-Swedish words (fyllek, inkluder, konsekwent), fragments (effektivitets, sammanfatt) and plural forms (distraktioner ‘distractions’, konditioner ‘conditions’, konflikter ‘conflicts’). At level 4, the model generates plausible words.

GPT-4o At level 0, the model generates fragments (gat), interjections (hej ‘hello’), but also some valid forms that are simply not in the reference word list (snart ‘soon’, idag ‘today’, snälla ‘please’). At level 1, the model generates plural forms that are mostly encountered in the plural (skor ‘shoes’, grönsaker ‘vegetables’, pengar ‘money’, byxor ‘pants’). At level 2, the model generates valid forms that are not present in the reference word list (plommon ‘plum’, citron ‘lemon’, fjärrkontroll ‘remote control’, körsbär ‘cherry’, fikon ‘fig’). At levels 3 and 4, the model generates plausible words.

Llama3-8B At level 0, the model generates personal pronouns (hon ‘she’, ni ‘you-PL’), conjunctions (om ‘if’), and multi-word expressions (du kan ‘you can’, ni är ‘you-PL are’, vi har ‘we have’). At level 1, the model generates plural forms (frukter ‘fruits’, händer ‘hands’, fötter ‘feet’), genitive forms (husdjurs ‘of the pet(s)’), and non-Swedish words (fartyk, probably meant to be fartyg ‘vehicle’). At level 2, the model generates plural forms (kängor ‘boots’, tänder ‘teeth’, fingrar ‘fingers’) and definite forms (landet ‘the land’). At level 3, the model generates plausible words. At level 4, the model generates plausible words but also quite some plural/definite/genitive forms.

Mistral-7B At level 0, the model generates non-Swedish forms (ananass, kokka, ingokt), geni-

tive forms (köks ‘of the kitchen’), numbers, interjections (hej ‘hello’), personal pronouns (du ‘you’), and some questionable forms such as *mannis(ka)* and *kvinn(a)* that cannot be decomposed as indicated in Swedish. The first word should be *människa* ‘human’, there is no such word as *männis*, and the second word should be *kvinna* ‘woman’, again there is no such word as *kvinn*. At level 1, the model generates plural forms (skor ‘shoes’, pengar ‘money’, kakor ‘cookies’, byxor ‘pants’), definite plural forms (äpplen ‘the apples’), and non-Swedish words (fräj). At level 2, the model generates clearly English words (autumn, january, march, august, winter), and the number one-hundred-eleven (hundraettioett). At level 3, the model generates non-Swedish words (hedervidy) and some misspelled words (heteronynym, ockupation, perssonlighet). At level 4, the model generates plausible words.

Phi3-4K At level 0, the model generates noun phrases with articles (en liten flicka ‘a small girl’, en liten hund ‘a small dog’, en liten fisk ‘a small fish’, *en liten hus ‘a small house’), multi-word expressions (jag har ‘I have’), definite forms (katten ‘the cat’), interjections (hej ‘hello’), articles (det ‘the’), comparative adjective forms (äldre ‘older’). At level 1, the model generates personal pronouns (du ‘you’, hon ‘she’, han ‘he’), nouns with article (en bilspår ‘a car track’, en bil ‘a car’), question particles (hur ‘how’) and small phrases (du/han/hon/jag/det är ‘you/he/she/I/it is/are’). At level 2, the model generates small phrases with *låt oss* ‘let’s’ (träffa ‘meet’, spela ‘play’, . . .). At level 3, the model generates definite forms (skolan ‘the school’, dörren ‘the door’, gatan ‘the street’). At level 4, the model generates multi-word expressions (framtidens utveckling ‘future development’, *kulturella identitet ‘cultural identity’).

4.4.5 Dutch

Gemma-2B At level 0, the model generates articles (het ‘the’) and personal pronouns (ik ‘I’, jullie ‘you’, hij/zij ‘he/she’). At level 1, the model generates questionable words (*esensieel, contextueel ‘contextual’, opwinding ‘excitement’, verenigt ‘united’, *genuinen, opdrachten ‘commands’, *overschrokken, onvoorspelbaar ‘unpredictable’). At level 2, the model also generates questionable words (oplossingen ‘solutions’, vervuld ‘fulfilled’, opwinding ‘excitement’, transformatie ‘transformation’, verhoogd ‘elevated’, liberaliseren ‘liberalize’, opvolging ‘succession’, multidimensionaal ‘multidimensional’) and English words (delicate, aromatic). At level 3, the model generates questionable words (*exceptieel). At level 4, the model generates plausible words.

GPT-4o At level 0, the model generates days of the week (woensdag ‘Wednesday’, vrijdag ‘Friday’) and multi-word expressions (dank je ‘thank you’). At level 1 and 2, the model generates words with the diminutive *-je* ending (broodje ‘bread-DIM’, koekje ‘cake-DIM/cookie’). At level 3 and 4, the model generates plausible words.

Llama3-8B At level 0, the model generates questionable words related to games (spelletjeskistje ‘game box’, speelkaart ‘playing card’, spelletjesdoos ‘game box’, spelletjesbox ‘game box’, spelletje ‘game’, spelletjspak ‘game pack’) and words with the diminutive *-je* ending (hondje ‘dog-DIM’, huisje ‘house-DIM’, katje ‘cat-DIM’, autootje ‘car-DIM’). At level 1, the model generates plausible words, but also days of the week, numbers, multi-word expressions and diminutive expressions (broertje ‘brother-DIM’, zusje ‘sister-DIM’). At level 2, the model generates more diminutive forms (liedje ‘song-DIM’, broertje ‘brother-DIM’, koekje ‘cake-DIM/cookie’, muziekje ‘music-DIM’, broodje ‘bread-DIM’, pakketje ‘package-DIM’, zusje ‘sister-DIM’, spelletje ‘game-DIM’, briefje ‘letter-DIM’). At level 3 and 4, the model generates plausible words.

Mistral-7B At level 0, the model generates days of the week (donderdag ‘Thursday’, woensdag ‘Wednesday’), numbers (vier ‘four’, vijf ‘five’), months of the year (augustus ‘August’, oktober ‘October’), personal pronouns (jullie ‘you’, hij ‘he’) and multi-word expressions (hoe zoekt u? ‘how do you search?’, met vriendelijke groet ‘yours sincerely’). At level 1, the model generates diminutive forms (vierkantje ‘square-DIM’, bankje ‘bench-DIM’, tabletje ‘tablet-DIM’, hakje ‘heel-DIM’, bootje ‘boat-DIM’, klusje ‘chore-DIM’). At level 2, the model also generates diminutive forms (appeltje ‘apple-DIM’, dagje ‘day-DIM’). At level 3 and 4, the model generates plausible words.

Phi3-4K At level 0, the model generates diminutive forms (appeltje ‘apple-DIM’). At level 1, the model generates plural forms (dieren ‘animals’, rozen ‘roses’), superlative adjective forms (oudste ‘oldest’), personal pronouns (ik ‘I’), conjugated verb forms (eet[7] ‘eats/eat-IMP’) and days of the week (maandag ‘Monday’). At level 2, the model

[7] In Afrikaans, *eet* is the infinitive form of the verb ‘to eat’

generates plural forms (boodschappen ‘groceries’, vrienden ‘friends’, autos ‘cars’). At level 3, the model generates all days of the week and *kledingstukken* ‘garments’. At level 4, the model generates multi-word expressions (regionale economie ‘regional economy’, sociale kwesties ‘social issues’).

4.4.6 General Remarks

Overall, we see a common pattern in the generated graded word lists, namely a propensity to generate personal pronouns (you, he, it), days of the week, months of the year, and numbers. All those categories were excluded from the reference word list based on part-of-speech filtering. A common motive also seems to be food and animals.

The models also tend to generate phrases rather than single words at times; phrases and multi-word expressions are undeniable useful for language learners, but the models do not adhere to the prompt.

In contrast to the grading task, which does not require models to output any language, the vocabulary generation tasks shows some shortcomings of the models when it comes to *producing* language other than English. This is noticeable for Spanish (Gemma-2B, Mistral-7B), French (GPT-4o, Llama3-8B, Mistral-7B, Phi3-4K), Swedish (Gemma-2B, Llama3-8B, Mistral-7B), and Dutch (Gemma-2B).

Finally, especially for English, all models generate words of the highest complexity when prompted for words of level 4. This may well be a phrasing problem in the prompt, as we explicitly state 4 as the *highest* level, albeit for language learners.

One general problem that we noticed is that if the word to assess is (or could be interpreted as) an English word, apparently mostly related to computer programming (by, blank, score, index, column, sample, type), the model fails to recognize the word to assess. We also notice that sometimes the models score outside of the given range (5,6,7), repeats the input prompt, or generates additional explanations even though it was asked not to. This is especially true for Gemma-7B.

5 Conclusion

In this paper, we presented experiments of using small versions of large generative language models out-of-the-box for (1) grading vocabulary lists and (2) generating graded vocabulary lists. Results show that while most of the models may only be targeted at English, they perform quite well cross-linguistically at the task of *grading* vocabulary. However, when it comes to *producing* vocabulary, the quality suffers.

One key finding is that GLMs that perform well on the task of grading vocabulary can be used to grade vocabulary items with low observed frequency. This use case uses the strength of graded word lists and GLMs for synergy effects.

We have also shown that using a numerical scale rather than the CEFR scale yields better results. This may be because the language models have not had enough contact with CEFR material to learn and “understand” what the different levels mean. A numerical scale may be more generalizable in this case.

To answer the hypothesis put forward at the beginning of the paper: “GLMs are effective at grading vocabulary”, we can conclude that all tested models exhibited some form of grading ability, although the predicted scores do not exactly match the textbook-derived scores, leading to low accuracy, precision and recall. However, when taking into account adjacent accuracy (the prediction is considered correct if it is at most one level from the target level), we can see values up to 99% (see Table 7 in the Appendix A).

When it comes to generating vocabulary from scratch, GLMs can be a starting point, although their potential for generating large graded vocabulary lists seem limited and needs further investigation. The inclusion of inflected forms (plural forms, conjugated verb forms) is undesired for most purposes.

One (maybe unsurprising) finding is that the much larger base-model GPT-4o performed best on average, indicating the larger GLMs may be more accurate in grading and generating vocabulary lists. However, Mistral-7B showed promising performance at second place and thus might be a viable free option.

Overall, while generative language models show promise in grading vocabulary across languages, continued research and development are needed to enhance their performance and applicability in language learning contexts.

In the hopes that the data may be of use to other researchers in the field, we make the data available at `https://github.com/daalft/cefrlex_llm`.

6 Future Work

We noticed that all models show a general tendency towards the middle levels. Previous research on feature-based classifiers shows that these classifiers tend to perform well on the extremes of the scale, and tend to mix up the middle levels (Pilán et al., 2016; Alfter and Volodina, 2018). Hence, we could potentially use feature-based classifiers to confidently identify items at the extremes of the scale, and GLMs to classify the middle levels.

Prompt engineering would also be a possible avenue for future work. A chain-of-thought prompt as used by Enomoto et al. (2024) may be more effective at eliciting not only a grade but also the decision process for arriving at that grade, allowing for greater transparency. As LLMs are sensitive to prompt formulation (Sclar et al., 2024), experimenting with different prompt wording may also prove beneficial.

Finally, it would be interesting to investigate how fine-tuning models impacts performance. We suggest a scenario where fine-tuning is done on one language family (e.g., Romance) and tested on a different language family (e.g., Germanic), to check for language-agnostic transferability of the graded vocabulary concept.

Limitations

In this work, we investigate only European languages, giving the work a strong Eurocentric focus. It would be beneficial to extend the investigation to more non-European languages.

In this work, we only tested small models. It is highly possible that the larger models may yield better results. However, such models also require significantly more power, both computational and financial.

Finally, we only generate up to 100 words for each level for each language. The generation limit of the GLM can be circumvented through a chat with history by repeatedly asking for *another* set of 100 words and passing the previously generated answer as history. Alternatively, the LLM can be prompted to generate *texts* of a certain proficiency level, based on which frequency-level information about words can be extracted, simulating a learner-oriented textbook (comprehension) corpus.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Qin Cai, Martin Cai, Caio César Teodoro Mendes, Weizhu Chen, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Yen-Chun Chen, Yi-Ling Chen, Parul Chopra, Xiyang Dai, Allie Del Giorno, Gustavo de Rosa, Matthew Dixon, Ronen Eldan, Victor Fragoso, Dan Iter, Mei Gao, Min Gao, Jianfeng Gao, Amit Garg, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Jamie Huynh, Mojan Javaheripi, Xin Jin, Piero Kauffmann, Nikos Karampatziakis, Dongwoo Kim, Mahoud Khademi, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Ce Liu, Mengchen Liu, Weishung Liu, Eric Lin, Zeqi Lin, Chong Luo, Piyush Madan, Matt Mazzola, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Xin Wang, Lijuan Wang, Chunyu Wang, Yu Wang, Rachel Ward, Guanhua Wang, Philipp Witte, Haiping Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Sonali Yadav, Fan Yang, Jianwei Yang, Ziyi Yang, Yifan Yang, Donghan Yu, Lu Yuan, Chengruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. *Preprint*, arXiv:2404.14219.

David Alfter. 2021. *Exploring natural language processing for single-word and multi-word lexical complexity from a second language learner perspective*. Ph.D. thesis, University of Gothenburg, Sweden.

David Alfter, Lars Borin, Ildikó Pilán, Therese Lindström Tiedemann, and Elena Volodina. 2019. Lärka: From Language Learning Platform to Infrastructure for Research on Language Learning. In *Selected papers from the CLARIN Annual Conference 2018*, pages 1–14. Linköping University Electronic Press.

David Alfter and Johannes Graën. 2019. Interconnecting lexical resources and word alignment: How do learners get on with particle verbs? In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 321–326.

David Alfter and Elena Volodina. 2018. Towards Single Word Lexical Complexity Prediction. In *Proceedings of the 13th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 79–88.

Yigal Attali, Andrew Runge, Geoffrey T LaFlair, Kevin Yancey, Sarah Goodwin, Yena Park, and Alina A Von Davier. 2022. The interactive reading task:

Transformer-based automatic item generation. *Frontiers in Artificial Intelligence*, 5:903077.

Drilon Avdiu, Vanessa Bui, Klára Ptacinová Klimci, et al. 2019. Predicting learner knowledge of individual words using machine learning. In *Proceedings of the 8th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2019), September 30, Turku Finland*, 164, pages 1–9. Linköping University Electronic Press.

Hamid Ayub. 2024. GPT-4o: Sucessor of GPT-4? https://hamidayub.medium.com/gpt-4o-successor-of-gpt-4-8207acf9104e. Accessed: June 12, 2024.

Anthony Baez and Horacio Saggion. 2023. LSLlama: Fine-tuned LLaMA for lexical simplification. In *Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability*, pages 102–108, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Stefano Bannò, Hari Krishna Vydana, Kate M Knill, and Mark JF Gales. 2024. Can GPT-4 do L2 analytic assessment? *arXiv preprint arXiv:2404.18557*.

Ummugul Bezirhan and Matthias von Davier. 2023. Automated Reading Passage Generation with OpenAI's Large Language Model. *Computers and Education: Artificial Intelligence*, 5:100161.

Ummugul Bezirhan and Matthias von Davier. 2023. Automated reading passage generation with OpenAI's large language model. *Computers and Education: Artificial Intelligence*, 5:100161.

B Boras, E Smolić, G Gledec, and T Jagušt. 2024. Exploring the Educational Potential of Generative AI: An Application for Spelling Practice. In *INTED2024 Proceedings*, pages 6945–6952. IATED.

Vaclav Brezina and Dana Gablasova. 2015. Is there a core general vocabulary? Introducing the new general service list. *Applied Linguistics*, 36(1):1–22.

Annette Capel. 2015. The English Vocabulary Profile. In Julia Harrison and Fiona Barker, editors, *English Profile in Practice*, chapter 2, pages 9–27. Cambridge University Press.

Council of Europe. 2018. Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion Volume with New Descriptors. Accessed 09.03.2019 from www.coe.int/lang-cefr.

Averil Coxhead. 1998. *An academic word list*, volume 18. School of Linguistics and Applied Language Studies.

Averil Coxhead. 2011. The academic word list 10 years on: Research and teaching implications. *Tesol Quarterly*, 45(2):355–362.

Edwin Creely. 2024. Exploring the Role of Generative AI in Enhancing Language Learning: Opportunities and Challenges. *International Journal of Changes in Education*.

Luise Dürlich and Thomas François. 2018. EFLLex: A Graded Lexical Resource for Learners of English as a Foreign Language. In *11th International Conference on Language Resources and Evaluation (LREC 2018)*.

Yo Ehara, Issei Sato, Hidekazu Oiwa, and Hiroshi Nakagawa. 2018. Mining Words in the Minds of Second Language Learners for Learner-specific Word Difficulty. *Journal of Information Processing*, 26:267–275.

Taisei Enomoto, Hwichan Kim, Tosho Hirasawa, Yoshinari Nagai, Ayako Sato, Kyotaro Nakajima, and Mamoru Komachi. 2024. TMU-HIT at MLSP 2024: How Well Can GPT-4 Tackle Multilingual Lexical Simplification? In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 590–598.

Chloe Farr. 2024. *Unmasking ChatGPT: The Challenges of Using Artificial Intelligence for Learning Vocabulary in English as an Additional Language*. Ph.D. thesis.

Thomas François and Barbara De Cock. 2018. ELELex: a CEFR-graded lexical resource for Spanish as a foreign language. In *PLIN Linguistic Day 2018: Technological innovation in language learning and teaching*.

Thomas François, Núria Gala, Patrick Watrin, and Cédrick Fairon. 2014. FLELex: a graded lexical resource for French foreign learners. In *LREC*, pages 3766–3773.

Thomas François, Elena Volodina, Ildikó Pilán, and Anaïs Tack. 2016. SVALex: a CEFR-graded Lexical Resource for Swedish Foreign and Second Language Learners. In *LREC*.

Núria Gala, Thomas François, and Cédrick Fairon. 2013. Towards a French lexicon with difficulty measures: NLP helping to bridge the gap between traditional dictionaries and specialized lexicons. *E-lexicography in the 21st century: thinking outside the paper., Tallin, Estonia*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy,

Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open Models Based on Gemini Research and Technology. *Preprint*, arXiv:2403.08295.

Robert Godwin-Jones. 2024. Distributed agency in second language learning and teaching through generative AI. *Preprint*, arXiv:2403.20216.

Johannes Graën, David Alfter, and Gerold Schneider. 2020. Using Multilingual Resources to Evaluate CEFRLex for Learner Applications. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 346–355, Marseille, France. European Language Resources Association.

Daniel R Isbell. 2017. Assessing C2 writing ability on the Certificate of English Language Proficiency: Rater and examinee age effects. *Assessing Writing*, 34:37–49.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *Preprint*, arXiv:2310.06825.

Abdelhak Kelious, Matthieu Constant, and Christophe Coeur. 2024. Complex Word Identification: A Comparative Study between ChatGPT and a Dedicated Model for This Task. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3645–3653.

Reza Kianian, Deyu Sun, Eric L. Crowell, and Edmund Tsui. 2024. The Use of Large Language Models to Generate Education Materials about Uveitis. *Ophthalmology Retina*, 8(2):195–201.

Adam Kilgarriff, Frieda Charalabopoulou, Maria Gavrilidou, Janne Bondi Johannessen, Saussan Khalil, Sofie Johansson Kokkinakis, Robert Lew, Serge Sharoff, Ravikiran Vadlapudi, and Elena Volodina. 2014. Corpus-based vocabulary lists for language learners for nine languages. *Language resources and evaluation*, 48(1):121–163.

Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, et al. 2023. ChatGPT: Jack of all trades, master of none. *Information Fusion*, 99:101861.

Cunliang Kong, Yun Chen, Hengyuan Zhang, Liner Yang, and Erhong Yang. 2022. Multitasking Framework for Unsupervised Simple Definition Generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5934–5943, Dublin, Ireland. Association for Computational Linguistics.

Richard LaBontee. 2019. *Strategic Vocabulary Learning in the Swedish Second Language Context*. Ph.D. thesis, University of Gothenburg.

Batia Laufer. 2006. Comparing focus on form and focus on forms in second-language vocabulary learning. *Canadian Modern Language Review*, 63(1):149–166.

John Lee and Chak Yan Yeung. 2018. Automatic prediction of vocabulary knowledge for learners of Chinese as a foreign language. In *2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP)*, pages 1–4. IEEE.

Joanne Leong, Pat Pataranutaporn, Valdemar Danry, Florian Perteneder, Yaoli Mao, and Pattie Maes. 2024. Putting Things into Context: Generative AI-Enabled Context Personalization for Vocabulary Learning Improves Learning Motivation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–15.

Jianhua Lin. 1991. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151.

Mounica Maddela and Wei Xu. 2018. A Word-Complexity Lexicon and A Neural Readability Ranking Model for Lexical Simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3749–3760.

I. Miralpeix and C. Muñoz. 2018. Receptive vocabulary size and its relationship to EFL language skills. *IRAL-International Review of Applied Linguistics in Language Teaching*, 56:1–24.

Kai North, Tharindu Ranasinghe, Matthew Shardlow, and Marcos Zampieri. 2023a. Deep learning approaches to lexical simplification: A survey. *arXiv preprint arXiv:2305.12000*.

Kai North, Marcos Zampieri, and Matthew Shardlow. 2023b. Lexical complexity prediction: An overview. *ACM Computing Surveys*, 55(9):1–42.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. GPT-4 Technical Report. *Preprint*, arXiv:2303.08774.

Jenny A Ortiz-Zambranoa and Arturo Montejo-Ráezb. 2020. Overview of ALexS 2020: First Workshop on Lexical Analysis at SEPLN.

Gustavo Paetzold and Lucia Specia. 2016. SemEval 2016 Task 11: Complex Word Identification. In *SemEval at NAACL-HLT*, pages 560–569.

Pearson. 2017. GSE Teacher Toolkit. https://www.english.com/gse/teacher-toolkit/user/vocabulary. Accessed: 2024-06-10.

Zhenhui Peng, Xingbo Wang, Qiushi Han, Junkai Zhu, Xiaojuan Ma, and Huamin Qu. 2023. Storyfier: Exploring Vocabulary Learning Support with Text Generation Models. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–16.

Ivica Pesovski, Ricardo Santos, Roberto Henriques, and Vladimir Trajkovik. 2024. Generative ai for customizable learning experiences. *Sustainability*, 16(7).

Ildikó Pilán, Elena Volodina, and Torsten Zesch. 2016. Predicting proficiency levels in learner writings by transferring a linguistic complexity model from expert-written coursebooks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2101–2111.

Ildikó Pilán, David Alfter, and Elena Volodina. 2016. Coursebook texts as a helping hand for classifying linguistic complexity in language learners' writings. In *Proceedings of the workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*. COLING 2016. Osaka, Japan.

Adrielli Rego Lopes, Joshua Snell, and Martijn Meeter. 2024. Language Models Outperform Cloze Predictability in a Cognitive Model of Reading.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting. *Preprint*, arXiv:2310.11324.

Matthew Shardlow. 2013. A Comparison of Techniques to Automatically Identify Complex Words. In *51st annual meeting of the association for computational linguistics proceedings of the student research workshop*, pages 103–109.

Matthew Shardlow, Fernando Alva-Manchego, Riza Batista-Navarro, Stefan Bott, Saul Calderon Ramirez, Rémi Cardon, Thomas François, Akio Hayakawa, Andrea Horbach, and Anna Huelsing. 2024a. The BEA 2024 Shared Task on the Multilingual Lexical Simplification Pipeline. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 571–589, Mexico City, Mexico. Association for Computational Linguistics.

Matthew Shardlow, Kai North, and Marcos Zampieri. 2024b. A Multilingual Survey of Recent Lexical Complexity Prediction Resources through the Recommendations of the Complex 2.0 Framework. In *Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context@ LREC-COLING 2024*, pages 51–59.

Philip Ernest Springer. 2012. *Advanced learner writing: A corpus-based study of the discourse competence of Dutch writers of English in the light of the C1/C2 levels of the CEFR*. Ph.D. thesis, University of Amsterdam, Netherlands.

Yao-Ting Sung, Wei-Chun Lin, Scott Benjamin Dyson, Kuo-En Chang, and Yu-Chia Chen. 2015. Leveling L2 texts through readability: Combining multilevel linguistic features with the CEFR. *The Modern Language Journal*, 99(2):371–391.

Anaïs Tack, Thomas François, Piet Desmet, and Cédrick Fairon. 2018. NT2Lex: A CEFR-Graded Lexical Resource for Dutch as a Foreign Language Linked to Open Dutch WordNet. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 137–146.

Anaïs Tack, Thomas François, Anne-Laure Ligozat, and Cédrick Fairon. 2016. Evaluating Lexical Simplification and Vocabulary Knowledge for Learners of French: Possibilities of Using the FLELex Resource. In *LREC*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *Preprint*, arXiv:2302.13971.

Yiwei Wang, Yujun Cai, Muhao Chen, Yuxuan Liang, and Bryan Hooi. 2023. Primacy Effect of ChatGPT. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 108–115.

Stuart Webb and Paul Nation. 2017. *How vocabulary is learned*. Oxford University Press.

Rodrigo Wilkens, David Alfter, Xiaoou Wang, Alice Pintard, Anaïs Tack, Kevin P Yancey, and Thomas François. 2022. Fabra: French aggregator-based readability assessment toolkit. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1217–1233.

Kevin Yancey and Yves Lepage. 2018. Korean L2 Vocabulary Prediction: Can a Large Annotated Corpus be Used to Train Better Models for Predicting Unknown Words? In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack, and Marcos Zampieri. 2018. A Report on the Complex Word Identification Shared Task 2018. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 66–78, New Orleans, United States. Association for Computational Linguistics.

Jiaxin Yuan, Cunliang Kong, Chenhui Xie, Liner Yang, and Erhong Yang. 2022. COMPILING: A Benchmark Dataset for Chinese Complexity Controllable Definition Generation. In *Proceedings of the 21st Chinese National Conference on Computational Linguistics*, pages 921–931, Nanchang, China. Chinese Information Processing Society of China.

Jiawei Zhao, Zhenyu Zhang, Beidi Chen, Zhangyang Wang, Anima Anandkumar, and Yuandong Tian. 2024. GaLore: Memory-Efficient LLM Training by Gradient Low-Rank Projection. *Preprint*, arXiv:2403.03507.

Imran Zualkernan and Salsabeel Shapsough. 2024. Towards Using Large Language Models to Automatically Generate Reading Comprehension Assessments for Early Grade Reading Assessment. In *INTED2024 Proceedings*, pages 3772–3782. IATED.

Anika Milena Żerkowska. 2024. Personalized Language Learning in the Age of AI: Leveraging Large Language Models for Optimal Learning Outcomes. Master's thesis.

A Generating grades: Full result table

	Gemma-2B					Gemma-7B					GPT-4o				
	Acc	AAcc	P	R	F1	Acc	AAcc	P	R	F1	Acc	AAcc	P	R	F1
EFLLex	0.20	0.84	0.21	0.20	0.18	0.24	0.93	0.32	0.24	0.18	**0.30**	**0.94**	**0.39**	**0.30**	**0.29**
ELELex	0.18	0.77	0.23	0.18	0.16	0.23	0.93	0.30	0.23	0.19	0.25	0.91	**0.33**	0.26	0.25
FLELex	0.21	0.85	0.36	0.22	0.15	0.17	0.89	0.39	0.18	0.12	0.21	0.91	0.36	0.22	0.22
SVALex	0.26	0.94	0.25	0.27	0.24	0.30	**0.96**	0.31	0.30	0.21	**0.33**	**0.96**	**0.34**	**0.34**	**0.33**
NT2Lex	0.24	0.90	0.30	0.24	0.18	0.27	0.97	0.28	0.27	0.19	0.33	0.97	0.39	0.34	0.35
	LLaMA3-8B					Mistral-7B					Phi3-4K				
	Acc	AAcc	P	R	F1	Acc	AAcc	P	R	F1	Acc	AAcc	P	R	F1
EFLLex	0.21	0.93	0.35	0.22	0.16	0.26	0.87	0.35	0.27	0.24	0.22	0.91	0.33	0.22	0.15
ELELex	0.24	**0.95**	0.29	0.24	0.20	**0.30**	0.90	0.31	**0.30**	**0.28**	0.25	0.94	0.35	0.25	0.19
FLELex	0.25	**0.96**	0.40	0.26	0.19	0.28	0.92	0.39	0.28	**0.28**	**0.29**	**0.96**	**0.42**	**0.29**	0.22
SVALex	0.28	**0.96**	0.31	0.29	0.25	0.21	0.87	0.28	0.21	0.18	0.26	0.95	0.28	0.26	0.20
NT2Lex	0.37	**0.99**	0.38	0.38	0.35	0.37	0.97	0.36	0.37	0.36	**0.40**	**0.99**	**0.40**	**0.40**	**0.38**

Table 7: Results in terms of Accuracy (Acc), Adjacent accuracy (AAcc), Precision (P), Recall (R), F1 score (F1), all weighted by label. Results in bold indicate the best result per category (Acc, AAcc, P, R, F1)

Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2024)

B Model parameters

Model parameters for generation. For the Gemma models and GPT-4o, no additional parameters were passed. For Mistral-7B and Llama3-8B, sampling was enabled, for Llama3-8B the `temperature` and `top_p` parameters were set, and for Phi-3, temperature was explicitly set to zero. The example code for Phi-3 additionally includes `do_sample=False`, which has no effect when `temperature` is zero, thus we excluded this parameter.

Model	Generation parameters
Gemma-2B	None
Gemma-7B	None
Mistral-7B	`do_sample=True`
Llama3-8B	`do_sample=True, temperature=0.6, top_p=0.9`
Phi-3	`temperature= 0.0`
GPT-4o	None

Table 8: Model generation parameters