Automated Evaluation of the Linguistic Difficulty of Short Texts for LLM Applications

Anonymous ACL submission

Abstract

There is an unmet need to evaluate the language difficulty of short passages of text, particularly for training and filtering Large Language Models (LLMs). Existing datasets fail to train models for this task, so we introduce ShortDiff, a new dataset with 890 short text passages in English together with their level of text difficulty. We experiment with a variety of models on ShortDiff, including finetuning Transformer-based models and prompting LLMs. Our best model achieves accuracy surpassing human experts and has latency appropriate to production environments. Finally, we release the ShortDiff dataset to the public for further research and development.

1 Introduction

In the domain of language acquisition tools, a key capability is the measurement of the linguistic difficulty of text. Traditionally, this has been used to assess a language learner's ability by evaluating their writing (Arnold et al., 2018; Ballier et al., 2019; Kerz et al., 2021). However, with the advent of use of Large Language Models (LLMs) for language practice and learning (Bonner et al., 2023; Kwon, 2023; Mahajan, 2022; Young and Shishido, 2023), a novel application has arisen: adjusting the language output of an LLM to the ability of a specific learner. The goal is to reduce the difficulty for beginners, and increase it for more advanced users, to maximize the user's learning by keeping them in the Zone of Proximal Development (ZPD) (Kinginger, 2002).

While LLMs have a degree of understanding of text complexity, this typically takes the form of text simplification, especially on large bodies of text (Espinosa-Zaragoza et al., 2023; Cardon and Bibal, 2023). In contrast, language learning requires exposure to short, authentic text segments (Leow, 1997), such as conversation. While LLMs are uniquely positioned to provide this, they are not typically trained to adjust short text output to the level of a learner.

To be able to make that adjustment, it is preferable to create an automated way to measure the linguistic difficulty of short passages of text. This methodology can then be used in an LLM-driven system to generate training data, annotate input contexts, and filter candidates, as depicted in example system diagram Figure 1. Critically, the text content must be analogous to the texts desired, which means short, preferably conversational passages.

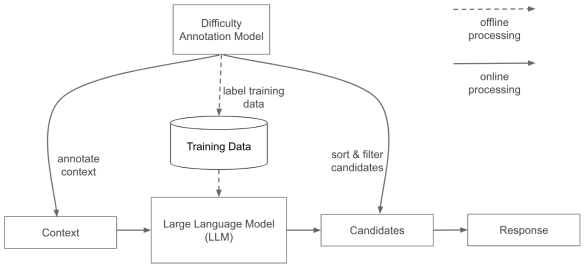

Figure 1: Example system diagram of LLM trained to produce text at different levels of difficulty, with a Difficulty Annotation Model required to label text at three points in the processing pipeline

While there is a significant body of work on the evaluation of text for difficulty, ranging from simple models like Flesch–Kincaid (Flesch, 2007) to techniques using neural networks (Filighera et al., 2019), these are unsuitable for handling short text passages, primarily because they are trained on long passages. (See Dataset section.)

What is needed is a new model that can accurately evaluate the difficulty of short, conversational pieces of text. Further, because such a model must be used both offline and online (see Figure 1), it must be fast enough to be in the critical path. In order to build such models, a suitable training and evaluation set is also required.

For research covered in this paper, we are focused on English language learning, but the tech-

niques should scale to other languages as well.

2 ShortDiff Dataset

There are a number of existing datasets related to measuring the text difficulty for second language (L2) acquisition of English. These include the English First Cambridge open language Database (EFCAMDAT) (Geertzen et al., 2014), the Cambridge Learner Corpus for the First Certificate in English (CLC-FCE) (by Lexical Computing Limited on behalf of Cambridge University Press and Assessment., 2017), and the Common European Framework of Reference for Languages (CEFR) (cef) leveled dataset provided by Adam Montgomerie (Montgomerie) (Montgomerie). However, these are unsuitable for training models to evaluate short, conversational passages of text, for several reasons.

First, these passages are too long, primarily because they are meant to establish a representative sample of a learner's abilities (Shatz, 2020). Even the passages in the shorter, non-evaluative Montgomerie set are over 400 words long on average, whereas the average turn length in a conversation is approximately 10 words (Yuan et al., 2006). Our experiments training a model on long passages and applying it to short ones proved ineffective, resulting in behavior that worked well only on passages of similar lengths. As an example, training a BERT-based model on the Montgomerie set and testing on our test set resulted in a Mean Squared Error of 1.81, double even the simplest Linear Model we tested, and almost 5 times larger than our best model.

Second, the most commonly-used datasets (EFCAMDAT and CLC-FCE) are comprised of examples authored by language learners. This makes them ideal for evaluating learners, but they are inappropriate for training LLMs to generate native-sounding speech. Finally, the distribution of difficulties is uneven, with especially few examples at high levels. This makes it difficult to train models capable of a wide range of evaluation.

Therefore, to train, evaluate, and compare our models, we created and labeled a novel dataset of short passages of text, in close collaboration with human language experts.

The ShortDiff dataset is comprised of 890 short text passages in English, created specifically for this task, split into training (445) and test (445). The average length of a passage is 12 words, with a median of 10. The shortest are 62 passages of a single word each and the longest passage is 114 words.

The provenance of the dataset is a mix of sources: generated internally for other language practice features (272), authored for this task by English language learning experts (255), generated by LLMs (198), anonymized segments from conversations with language learners (101), and public data from the web (64). Much of the dataset is selected to be conversational in nature, since that is the primary expected application.

The dataset was labeled in batches of approximately 100, with sampling adjusted with the goal of an approximately even distribution along the CEFR scale, to include a range of beginner, intermediate, and advanced texts. While C1, C2, and A1 texts are slightly underrepresented, subsampling can be applied to get an even distribution if desired (Figure 2).

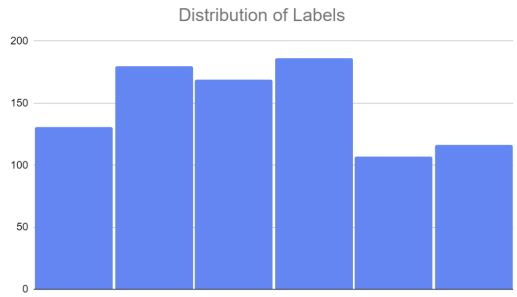

Figure 2: Distribution of CEFR levels in the ShortDiff dataset, as labeled by human expert raters. The distribution of floor(label) is A1: 131, A2/A2+: 180, B1/B1+: 169, B2/B2+: 186, C1: 107, C2: 116)

For C1 and C2 levels, language experts created examples using both advanced vocabulary (e.g., "He feigned indifference.") and colloquial and idiomatic usage (e.g., "Get off your high horse and lend me a hand. This house isn't going to paint itself.")

2.1 Human Expert Labels

Passages in the dataset were rated by English language learning experts (each with at least a Master's degree in Applied Linguistics or similar, plus a minimum of 10 years of experience in language teaching, language teaching curricula and assessment development, teacher education, or research in the field). Labels were applied on the CEFR scale (cef): A1 through C2. By convention, the labels A2 through B2 include "+" variations, indicating a level higher than the baseline.

Each passage was labeled by at least two raters, working independently, but collaborating on a rating guideline document to align themselves. The CEFR labels were applied based on the productive difficulty, i.e., the level at which an L2 learner can be expected to produce the text. When labeling single words, the meaning with the lowest level was chosen, as that is most likely to be used by a language learner.

Ratings were then converted to numbers (A1=1, A2=2, A2+=2.5, B1=3, B1+=3.5, B2=4, B2+=4.5, C1=5, C2=6), and averaged to arrive at a consensus per passage. In some cases, more raters were available and we included those in the average (112 cases). In about 5% of cases, due to differences greater than 1 between individual raters, labels were adjudicated by expert raters as a group to arrive at a consensus label. At the end of model training, the worst 20 predictions from each model on the test set were re-adjudicated to identify potential mislabels (123 cases of adjudication total).

To compare our models against an unbiased metric, one more set of ratings was performed, just on the test set, by an expert who did not previously work with other raters, but who used the rating guideline as well as training set labels for calibration.

3 Evaluation Framework

We evaluated our models on predicting the labels in the human-rated test set. Because of averaging between raters, the labels are not constrained to exact CEFR boundaries, e.g., "I have lived here since I was 4." is labeled 2.75, meaning that it falls between the A2+ and B1 CEFR labels. Our primary metric was therefore chosen to be Mean Squared Error (MSE) between a model's predictions and the consensus human expert label, on the 1-6 scale, meaning the maximum error possible is 5, and accordingly the maximum MSE is 25.

The independent expert human rater who did not work with the original raters achieved a MSE of 0.75 (90% confidence interval [0.67, 0.84]). For additional reference, we also evaluated the original primary raters who collaborated on the dataset labels. They were measured against the average of all ratings other than their own (including the independent rater), or the adjudicated label if there was one. They had MSEs of 0.47 ([0.41, 0.53]) and 0.54 ([0.48, 0.61]). However, since they worked closely together and collaborated on adjudication, this is a biased comparison point.

While most human expert disagreements were within 1 point of one another, 8% of the labels were further apart than this. Disagreements were particular common for intermediate CEFR levels (Figure 3).

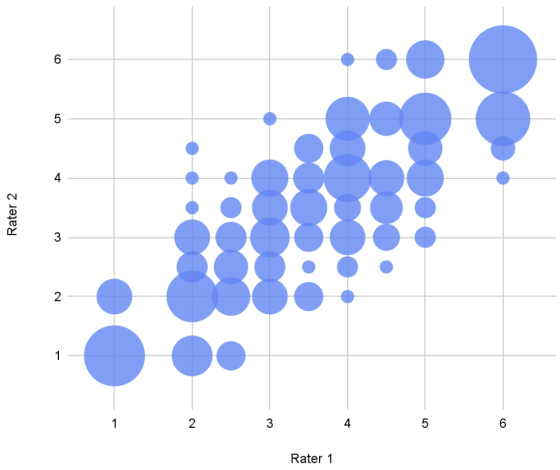

Figure 3: Label agreement between the two primary expert raters. Circle sizes represent the number of passages with each pair of labels. Significantly more disagreement occurs toward the middle of the CEFR scale than at each end.

We took the independent expert labeler MSE of 0.75 as the main target for machine earning models, although ultimately we were able to surpass the biased metrics of the primary raters as well.

4 Models Overview

We evaluated three types of models, in order from simplest to most complex: a linear regression model on surface language features, a custom model fine-tuned off Bidirectional Encoder Representations from Transformers (BERT), and a Large Language Model (PaLM 2-L) (Anil et al., 2023) in a few-shot setting. Summary of results is in Figure 4 and Table 1.

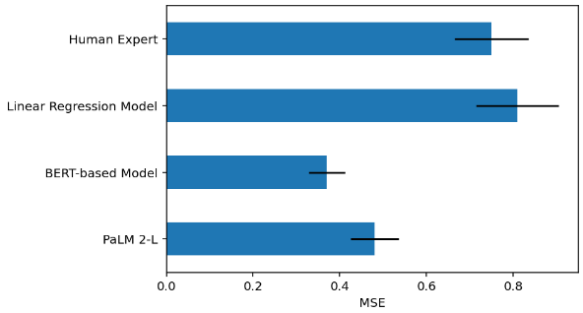

Figure 4: Mean Squared Error for different model types, with 90% confidence intervals.

Table 1: Accuracy Summary

Model Type	MSE	Correlation to Label
Human Expert	0.75	0.88
Linear Model on Surface Features	0.81	0.81
BERT-based Model	0.37	0.92
PaLM 2-L	0.48	0.9

In addition to accuracy, latency is critical for practical consideration. Some use cases, like generating offline training data, are relatively latency insensitive, but others are in the critical path, like integrating with an LLM for generation (Figure 1) or evaluating user proficiency in real time. This means for key applications, a model with latency in the 10ms to 100ms is necessary. Latency results summary is in Table 2.

Table 2: Latency summary. Latency is extremely approximate, and no effort has been made to optimize for speed. Note further that GPU and TPU execution is highly parallelizable, so amortized batch lookup speed is significantly faster than individual lookup)

Model Type	Latency CPU (One lookup)	Latency GPU/TPU (One lookup)
Linear Model on Surface Features	~50μs	-
BERT-based Model	~100ms	~10ms
PaLM 2-L	-	~1s

5 Linear Regression Model

The linear regression model is a simple algebraic model optimizing for the label from surface characteristics of text like average sentence and word lengths.

The benefit of such models is their simplicity and speed. The model we built can execute locally in-process, with latency measured in microseconds. The downside is that their accuracy is extremely limited because of a lack of understanding the text in any way.

5.1 Features

There is considerable prior research on measuring text difficulty, using surface features such as sentence and word length (Khushik and Huhta, 2022) or word diversity (Treffers-Daller et al., 2018). While these are not encompassing metrics of text complexity (Tanprasert and Kauchak, 2021), they correlate strongly with difficulty. After experimentation, we settled on the signals "average word length in characters," "average sentence length in characters," and "average sentence length in words" (Figure 5).

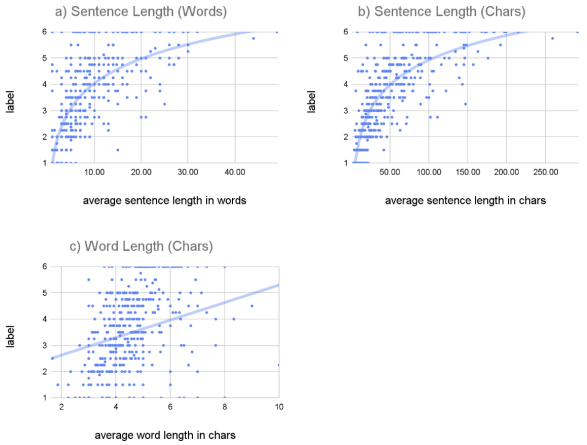

Figure 5: Correlation between linear model signals and label on train set. Correlations are 0.67, 0.70 and 0.35 for average sentence length in words, average sentence length in chars and average word length in chars respectively. Notably the sentence length signal has a logarithmic relationship to the signal, and correcting for that by taking ln(signal) improves the correlations to 0.71 and 0.75 for those signals respectively.

The key weakness of these features is they are content agnostic. For example, "The cat is here." (A1 difficulty) and "The apex of ire." (C1/C2 difficulty) have indistinguishable word and sentence features. For these reasons, such approaches are most effective when averaged over long texts, and suffer greatly from the brevity of our dataset.

5.2 Results

Of the models tested, the linear model performed the worst (Figure 4), with an MSE of 0.81 (90% confidence [0.71-0.91]). Typical errors relate to mistaking the difficulty of short words and sentences comprised of them (Table 4). It also tends to overestimate the difficulty of sentences that are

simple in structure but have many words, e.g., "For herbal tea, we have blueberry chamomile, chai, rooibos, fennel tarragon, and nettle." is labeled at 3 (B1) but predicted by the model to be 5 (C1)

6 Large Language Model

An LLM is a natural choice for evaluating the difficulty of text. Such models have intrinsic understanding of language, and their training data often organically include the CEFR scale (Yancey et al., 2023). It is possible to ask an LLM to evaluate a passage of text and get a reasonable response. The downside is that these models are comparatively slow (Table 2) and are therefore primarily suitable for offline text labeling.

We used the PaLM 2-L model (Anil et al., 2023), a model optimized for language understanding, generation, and translation tasks. We limited ourselves to few-shot prompt engineering. It is likely that prompt tuning or fine tuning would yield better results, and this is a direction where further research is ongoing.

6.1 Results

6.1.1 Initial Results

For the initial results, we used a single prompt, populated by instructions and examples from the training data. Notably, because of the constraints of context length, we randomly sampled 64 out of 445 training examples. This resulted in an MSE of 0.98.

6.1.2 Averaging Across Training Data

Since the limitation of the context length prevented us from using all of the training data, we experimented with running the model multiple times, re-sampling the training data, and averaging the results. By rerunning the model 3 times, we improved accuracy, from an MSE of 0.98 to 0.78. Naturally, this results in proportionately increased latency. Further improvement is likely possible if more samples are taken.

6.1.3 Splitting out Individual Words

We noted that the model had significant difficulty predicting the label of single words compared to phrases. We hypothesized that this is because from the LLM's perspective, these are very different tasks, and because many more of the training examples are phrases (N=418) compared to single words (N=27). Since the training examples are further subsampled in sets of 64 to fit in the context, only 3-4 single words would actually be seen by the model.

To address this, we separated the prompts into two types: one responsible for predicting the difficulty of phrases, and another one for predicting the difficulty of individual words (Appendix A) This significantly improved the MSE, from 0.78 to 0.48.

6.1.4 Final Results

The final results are an MSE of 0.48 (90% confidence [0.43, 0.54]) (Chart 4). This is substantially better than the linear model, and much better than human expert ratings, albeit at a significant latency cost (Table 2). Unlike the linear model, there's no obvious pattern of errors (Table 5). The opacity of mistakes is a risk factor, since this can make it challenging to improve the model further.

7 BERT-based Model

The BERT-based model builds on an existing, lightweight BERT encoder, which provides a combination of a high degree of accuracy and production-level latency. We fine-tuned a custom model by taking the first few layers of pretrained BERT model and adding a classification head. The BERT encoder is multiple orders of magnitude smaller than a typical LLM (millions rather than billions of parameters), but still comes pretrained with a degree of language understanding, and is fine-tunable to very specific tasks. It is also well-suited to serve as a distilled version of a larger model, which we used during quality iteration.

7.1 Results

7.1.1 Initial Results

We fine-tuned the BERT encoder on the 445 training samples. We ran light hyperparameter tuning (on a validation set split from the training samples) for the number of layers of the pretrained encoder to keep, learning rate, batch size, and warm up proportion. BERT achieved an MSE of about 0.44, which is substantially better than any of the other models.

7.1.2 2-Stage Finetuning with LLM Labeling

Unlike the linear model, which peaks in accuracy after a few dozen examples, and the LLM, which is context-constrained to accept only a few dozen examples, the BERT model continues to improve with additional training data. We therefore added an extra finetuning stage to the training. In the first

stage, we labeled 10 thousand examples from various sources with our best LLM version. We used those LLM-labeled examples to finetune the BERT model. In the second stage, we further finetuned the model on the human expert rated dataset. The results improved significantly, from MSE 0.44 to 0.37.

7.1.3 Final Results

The final results are an MSE of 0.37 (90% confidence [0.32, 0.41)] (Chart 4), which is a dramatic improvement over human experts and the other models. The latency, particularly when running on GPU (Table 2) is also practical enough for latency-sensitive production applications, making this the ideal model for most use cases.

The only recurring issue we saw was that this model struggled with misspellings, compared to the LLM (with its larger vocabulary) and the Linear Model (which has no concept of spelling). We did not deliberately introduce misspellings into the ShortDiff dataset, but they arose naturally from several of our sources. Ultimately, we decided to correct the spellings, because we want to be able to also use the dataset for generative tuning, and don't want to train models to produce misspellings. However, this is a weakness that needs to be taken into account when integrating into production use cases, and a spell-checker may be helpful.

Aside from misspellings, the BERT-based model's errors were similarly opaque to the LLM errors. The only significant pattern was having difficulty with idiomatic sayings like "It's been a rough spell but I'm game to try anything that might help us weather this storm." (Table 6)

8 Ensemble Models

It is noteworthy that while each model makes mistakes, the categories of mistakes made by different models differ. This makes sense, since, for example, the Linear Model has no concept of language meaning, whereas the BERT model has no concept of word length. We therefore evaluated whether it's possible to offset the errors of the different models by combining them together.

To do so, we randomly split out 100 examples from the test set to use for tuning, and used the remaining 355 examples for evaluation. We weighted the models to optimize performance on the tuning set, essentially putting a linear model over them. With this approach, we were able to reduce MSE from 0.36 for BERT to 0.33 when combining BERT+LLM. Adding the linear model to the mix did not improve results further beyond noise levels. Figure 6.

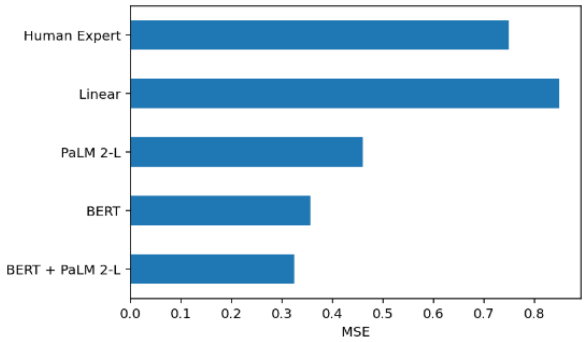

Figure 6: MSE of ensemble models.

While this improvement is incremental, and likely incurs too much complexity to be used in production, it is helpful for establishing that further improvements in accuracy are possible, and this approach may be useful for creating better pre-training datasets for improvements to BERT in the future.

9 Summary

Ultimately, we were able to achieve accuracy better than expert human ratings on short conversational pieces of text. We are releasing the ShortDiff dataset to the public for further iteration, and have been successfully integrating the models into LLM systems designed to help learners practice in an authentic conversational setting.

10 Limitations

The ShortDiff dataset provides the ability to train models on short pieces of text, but it still has several limitations. It was generated from a limited set of sources, and rated by a small cohort of expert raters. Diversifying both the sources and the raters may provide significantly less biased and more generalized results. Additionally, the dataset and all the models trained on it here are limited to English, which does not serve populations trying to learn other languages. Expanding the dataset to other languages is possible, but would require incremental work per language unless an automated methodology is identified.

Another significant limitation of these approaches is that they rely on a single scale for difficulty, which is not representative of the diverse experiences and backgrounds of learners. A

more fine-grained and personalized approach to user challenge is going to be made possible by the advent of LLMs, and is a fertile ground for future research.

11 Future Work

The next natural step is integrating this work into LLM generation, using both the manually labeled difficulty dataset and the automated difficulty measuring models.

Additionally, there is considerable work to be done to improve the dataset, as mentioned in the Limitations section, including size, diversity, and scaling to non-English languages.

Beyond that, there's still headroom to further improve accuracy, as demonstrated by the ensemble model experimentation. We believe that adding a dictionary of average word frequency or difficulty to the Linear model, such as the Global Scale of English dictionary (GSE) would significantly improve its results without sacrificing latency, though it's not expected it would surpass the language models. Such a dictionary could also be automatically generated using the larger models. Further work in distillation is also of great practical interest, particularly distilling LLM and BERT-based models into smaller versions with lower latency and operational costs.

References

Common european framework of reference for languages (cefr). Online.

Global scale of english. Online.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. Palm 2 technical report.

Taylor Arnold, Nicolas Ballier, Thomas Gaillat, and Paula Lissòn. 2018. Predicting cefrl levels in learner english on the basis of metrics and full texts.

Nicolas Ballier, Thomas Gaillat, Andrew Simpkin, Bernardo Stearns, Manon Bouyé, and Manel Zarrouk. 2019. A supervised learning model for the automatic assessment of language levels based on learner errors. In *Transforming Learning with Meaningful Technologies*, pages 308–320, Cham. Springer International Publishing.

Euan Bonner, Ryan Lege, and Erin Frazier. 2023. Large language model-based artificial intelligence in the language classroom: Practical ideas for teaching. *Teaching English with Technology*, 23(1).

Distributed by Lexical Computing Limited on behalf of Cambridge University Press and Cambridge English Language Assessment. 2017. Openclc (v1).

Rémi Cardon and Adrien Bibal. 2023. On operations in automatic text simplification. In *Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability*, pages 116–130.

Isabel Espinosa-Zaragoza, José Abreu-Salas, Elena Lloret, Paloma Moreda, and Manuel Palomar. 2023. A review of research-based automatic text simplification tools. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 321–330, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Anna Filighera, Tim Steuer, and Christoph Rensing. 2019. Automatic text difficulty estimation using embeddings and neural networks. In *Transforming Learning with Meaningful Technologies: 14th European Conference on Technology Enhanced Learning, EC-TEL 2019, Delft, The Netherlands, September 16–19, 2019, Proceedings 14*, pages 335–348. Springer.

Rudolf Flesch. 2007. Flesch-kincaid readability test. *Retrieved October*, 26(3):2007.

Jeroen Geertzen, Theodora Alexopoulou, and Anna Korhonen. 2014. Automatic linguistic annotation oflarge scale l2 databases: The ef-cambridge open language database(efcamdat).

Elma Kerz, Daniel Wiechmann, Yu Qiao, Emma Tseng, and Marcus Ströbel. 2021. Automated classification of written proficiency levels on the CEFR-scale through complexity contours and RNNs. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 199–209, Online. Association for Computational Linguistics.

Ghulam Abbas Khushik and Ari Huhta. 2022. Syntactic complexity in finnish-background efl learners' writing at cefr levels a1–b2. *European Journal of Applied Linguistics*, 10(1):142–184.

Celeste Kinginger. 2002. Defining the zone of proximal development in us foreign language education. *Applied linguistics*, 23(2):240–261.

Taeahn Kwon. 2023. *Interfaces for Personalized Language Learning with Generative Language Models*. Ph.D. thesis, Columbia University.

Ronald P Leow. 1997. The effects of input enhancement and text length on. *Applied Language Learning*, 8(2):151–182.

Muskan Mahajan. 2022. BELA: Bot for English language acquisition. In *Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI)*, pages 142–148, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Adam Montgomerie. Attempting to predict the cefr level of english texts. Online.

Itamar Shatz. 2020. Refining and modifying the efcamdat: Lessons from creating a new corpus from an existing large-scale english learner language database. *International Journal of Learner Corpus Research*, 6(2):220–236.

Teerapaun Tanprasert and David Kauchak. 2021. Flesch-kincaid is not a text simplification evaluation metric. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 1–14.

Jeanine Treffers-Daller, Patrick Parslow, and Shirley Williams. 2018. Back to basics: How measures of lexical diversity can help discriminate between cefr levels. *Applied Linguistics*, 39(3):302–327.

Kevin P Yancey, Geoffrey Laflair, Anthony Verardi, and Jill Burstein. 2023. Rating short l2 essays on the cefr scale with gpt-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 576–584.

Julio Christian Young and Makoto Shishido. 2023. Investigating openai's chatgpt potentials in generating chatbot's dialogue for english as a foreign language learning. *International Journal of Advanced Computer Science and Applications*, 14(6).

Jiahong Yuan, Mark Liberman, and Christopher Cieri. 2006. Towards an integrated understanding of speaking rate in conversation. In *Ninth International Conference on Spoken Language Processing*.

A LLM Prompts

Listing 1: Prompt to Evaluate Text Difficulty for Phrases

```
CEFR is a six-level scale, with each level
    ↪ corresponding to a specific level of English
    ↪  language proficiency. The levels are:

- A1 (1): Beginner
- A2 (2): Elementary
- B1 (3): Intermediate
- B2 (4): Upper Intermediate
- C1 (5): Advanced
- C2 (6): Proficiency

According to the CEFR scale, the proficiency level
    ↪ required to use the following phrases are:

Phrase: You are welcome! -> CEFR: 1
Phrase: I wonder if there's any treasure. -> CEFR:
    ↪ 3.25
[more examples...]
Phrase: {test_phrase} -> CEFR:
```

Listing 2: Prompt to Evaluate Text Difficulty for Single Words

```
GSE is a six-level scale, with each level
    ↪ corresponding to a specific level of English
    ↪  language proficiency. The levels are:

- A1 (1): Beginner
- A2 (2): Elementary
- B1 (3): Intermediate
- B2 (4): Upper Intermediate
- C1 (5): Advanced
- C2 (6): Proficiency

According to the GSE scale, the proficiency level
    ↪ required to use the following words are:

age,1
almost,2
[more examples...]
{test_word},
```

B Example Errors

Tables with the worst error examples from each model type.

Table 3: **Human Expert Rater**: worst 5 errors, labels are 1-6 with 1 corresponding to A1 on the CEFR scale and 6 corresponding to C2

Text	Label	Prediction	Error
The Sumida River is one of Japan's biggest, and you can take a tour on a boat and see the sights along the river's edges like sumida aquarium, temples, and more. The Sumida Observatory lets you take in a birdseye view of the river and Tokyo. Are you ready to book your tickets?	5	2.5	-2.5
I have a nice garden with flowers, trees, and a small pond.	3.25	1	-2.25
I like the classics over remakes.	4.75	2.5	-2.25
I see. Dulce de leche is a popular dessert in Argentina, and it is often used as a filling for pastries and other desserts. Empanadas are also a popular dish in Argentina, and they can be filled with a variety of ingredients, such as meat, cheese, or vegetables.	5.25	3	-2.25
I'm looking to the future with hope.	4.25	2	-2.25

Table 4: **Linear Model**: worst 5 errors, labels are 1-6 with 1 corresponding to A1 on the CEFR scale and 6 corresponding to C2

Text	Label	Prediction	Error
to ascertain	6	2.4	-3.6
naive	4	1.1	-2.9
endeavor	5	2.4	-2.6
Get off your high horse and lend me a hand. This house isn't going to paint itself.	6	3.6	-2.4
effervescent	6	3.6	-2.4

Table 5: **PaLM 2-L**: worst 5 errors, labels are 1-6 with 1 corresponding to A1 on the CEFR scale and 6 corresponding to C2

Text	Label	Prediction	Error
By perseverance.	4	1	-3
Just a couple of weeks.	1	3	2
By perseverance, just not giving up even when things seem impossible.	5.5	3.87	-1.63
The rate at which kids absorb new information is simply astonishing.	6	4.4	-1.6
Yeah, it's quite a controversy!	4.75	3.2	-1.55

Table 6: **BERT-based model**: worst 5 errors, labels are 1-6 with 1 corresponding to A1 on the CEFR scale and 6 corresponding to C2

Text	Label	Prediction	Error
hobby	1	3.23	2.23
Celery is a low calorie vegetable.	4	2.13	-1.87
I didn't understand the noise last night.	2.25	3.82	1.57
I am definitely leaning towards accepting it.	3.5	5.02	1.52
Get off your high horse and lend me a hand. This house isn't going to paint itself.	6.0	4.55	-1.45