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

Anonymous ACL submission

Abstract

There is an unmet need to evaluate the language difficulty of short, conversational passages of text, particularly for training and filtering Large Language Models (LLMs). We introduce Ace-CEFR, a novel dataset comprising 890 English conversational text passages, each annotated with its corresponding level of text difficulty. We experiment with a variety of models on Ace-CEFR, 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 Ace-CEFR 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 learning and practice (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 maximize the user's learning by keeping them in the Zone of Proximal Development (ZPD) (Kinginger, 2002), reducing the difficulty for beginners and increasing it for more advanced users .

While LLMs have a degree of understanding of text complexity, this typically takes the form of text simplification, especially on long text passages (Cardon and Bibal, 2023; Espinosa-Zaragoza et al., 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.

In order to make that adjustment, it is preferable to create an automated way to measure the linguistic difficulty of short, conversational passages of text. This can be used in an LLM-driven system to generate responses at a specific difficulty level. In this kind of system, a difficulty model can be applied at several points. The first is labeling training or fine-tuning data. The second is annotating the LLM prompt with difficulty labels for few-shot prompt engineering. The third is applying the difficulty model to the LLM output candidates to select the ones closest to the desired difficulty. An example system of this kind is shown in Figure 1. It is notable that these applications are a mix of offline and online processing, with the latter being highly sensitive to latency.

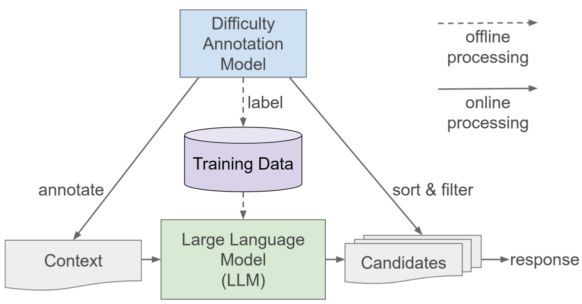

Figure 1: Example system diagram of LLM trained to produce text at different levels of difficulty, with a Difficulty Annotation Model required to label text at three points in the processing pipeline.

To be effective in this kind of system, the difficulty annotation model must be trained on text analogous to those the LLM is generating, which means short, conversational passages.

1.1 Summary of Contributions

The goal of this work is to author a new dataset, identify baselines for model performance on it, and establish that it's possible to train models applicable to practical, real-time applications. Our contributions are listed below.

- We release a new dataset, Ace-CEFR, for evaluating English language difficulty. The dataset can be used to train models to understand the difficulty of text, as well as to train LLMs to generate text at specified levels, or for related tasks such as complex word identification.

- We establish baselines for performance on the difficulty evaluation task, for both human experts and machine models of different levels of complexity.

- We demonstrate that it is feasible for relatively small models (a few million parameters) to achieve good accuracy on this task, with low latency, suitable for real-time applications.

1.2 Related Work

1.2.1 Datasets

There are a number of difficulty-annotated datasets at the document level, on the order of hundreds of words in length. 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), Weebit (Rama and Vajjala, 2021), OneStopEnglish (Vajjala and Lučić, 2018), Newsela (Nushi and Fadaei, 2020), who annotated passages with various readability measures, a dataset provided by Adam Montgomerie (Montgomerie, 2021) labeled on the CEFR scale, Wiki-Auto (Jiang et al., 2020), and the Sentence Corpus of Remedial English (SCoRE) (Chujo et al., 2015). In many cases, these texts are deliberately long to establish a representative sample of a learner's abilities (Shatz, 2020).

However, these are too long to train LLMs to produce conversational responses, being hundreds or more words long, compared to the average turn length in a conversation which is approximately 10 words (Yuan et al., 2006). We further cannot simply split the passages up and train models on subsections, because while some studies presumed the same readability for sentences within a document (Collins-Thompson and Callan, 2004; Dell'Orletta et al., 2011; Vajjala and Meurers, 2014; Ambati et al., 2016), this assumption has been shown to not hold (Arase et al., 2022).

There are a smaller number of datasets annotated at the sentence level. These include Štajner et al. (2017), which employed a 5-level scale to evaluate the complexity of human-written and machine-generated sentences, Brunato et al. (2018), who used a 7-level scale for sentences from news articles in linguistic databases (McDonald et al., 2013), and the CEFR-SP dataset (Arase et al., 2022) which contains English sentences annotated on the Common European Framework of Reference (CEFR) scale.

These shorter datasets are more closely aligned to our needs, but are still challenging to use directly for LLM training. The biggest obstacle is that they are not representative of conversations. The closest to our needs is the CEFR-SP dataset, but its passages are composed of uniform, single-sentence, complete-thought sentences, and do not include the variations typically seen in conversations such as phrases, single word responses, references to other parts of the conversation, or multiple sentences.

Further difficulties in training models on these datasets arise from unbalanced distributions of difficulties. The datasets are typically taken either from examples authored by language learners (e.g. EFCAMDAT and CLC-FCE), or sampled from natural text (e.g. CEFR-SP). This results in distributions that are highly skewed either toward the beginner levels or toward the middle of the difficulty curve, with almost no examples at high levels. This makes it difficult to train models capable of a wide range of evaluation. It is worth noting that, while examples authored by language learners are ideal for evaluating learners, they are inappropriate for training LLMs to generate native-sounding speech.

For these reasons, we decided to author and annotate a novel dataset, composed deliberately of short, conversational texts at a variety of levels, including single words, phrases, sentences, and short passages.

1.2.2 Modeling

A variety of automated models have been used for the evaluation of text difficulty, typically focusing on either readability, or alignment with the Common European Framework of Reference ((CEFR)) scale, a standardized measure of language difficulty for L2 learners.

Readability is a metric that tries to approximate how easy text is to read. There are multiple defined metrics (Matricciani, 2023) generally focused on the length and complexity of sentences and words. Readability of text has traditionally been estimated by combining word length or word frequency statistics with scaled sentence length (Stenner et al., 1988; Fry, 1990; Chall and Dale, 1995), Petersen and Ostendorf (2009). More recent works show that neural network-based approaches outperform statistical feature-based methods (Azpiazu and Pera, 2019; Meng et al., 2020; Imperial, 2021), (Martinc et al., 2021). Related efforts have focused on the word complexity aspect of readability specifically (Aleksandrova and Pouliot, 2023) (North et al., 2023).

However, readability is only representative of one aspect of difficulty, and many research efforts focus on the CEFR scale, which evaluates multiple dimensions of difficulty, especially for L2 learners. Salamoura and Saville (2010); Ishii and Tono (2018) explored aligning English vocabulary and grammar with CEFR levels. Uchida and Negishi (2018) experimented with automated CEFR level assessment at the passage level, using data from Cambridge English exams. Notably, Rama and Vajjala (2021) showcased the high accuracy of Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2018) in multilingual CEFR-level classification tasks, and Arase et al. (2022) developed a text CEFR level assessment model with BERT embeddings that performs significantly better than models based on superficial text features.

In alignment with these efforts, we have focused our modeling on the CEFR scale, but applied specifically our Ace-CEFR dataset. To establish a clear baseline for further work, we evaluated a representative range of models, including statistical feature engineering, neural networks, and LLM prompting, analyzing their respective characteristics.

2 Ace-CEFR Dataset

To address the lack of short, conversation datasets described in Section 1.2, we created a new dataset that draws from a diverse mix of sources, targeting conversational texts, and labeled them in close collaboration with human language experts.

The Ace-CEFR (Annotated ConvErsational CEFR-aligned) 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, aligned with typical conversation turn length (Yuan et al., 2006). There are 62 passages composed of a single word each, and the longest passage is 114 words.

The provenance of the dataset is a mix of sources: generated by our research organization for other language practice efforts (272), authored for the task of difficulty labeling by English language learning experts (255), generated by LLMs (198), anonymized segments from conversations with trusted tester language learners (101), and public data from the web (64). Anonymized conversation segments were processed via automated tools to remove potentially identifying information, and then further manually inspected and rewritten to ensure privacy. Much of the dataset is selected to be conversational in nature, since that is the primary expected application.

The texts were labeled aligned with the Common European Framework of Reference (CEFR) scale, a standard that organizes proficiency into six levels: A1-A2 (beginner), B1-B2 (intermediate), and C1-C2. In order to include examples of all levels, the dataset was labeled in batches of around 100, with a sampling method adjusted with the goal of a uniform distribution of levels. Although texts at the C1, C2, and A1 levels are somewhat underrepresented, subsampling techniques can be utilized to achieve a more balanced distribution if needed (Figure 2).

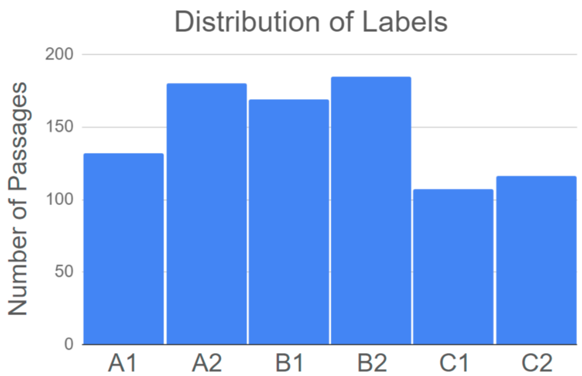

Figure 2: Distribution of CEFR levels in the Ace-CEFR dataset, as labeled by human expert raters. The distribution of floor(label) is A1: 131, A2/A2+: 180, B1/B1+: 169, B2/B2+: 186, C1: 107, C2: 116.

For the 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 (CEFR): A1 through C2. By convention, the labels A2 through B2 include "+" variations, indicating a level higher than the baseline.

Each text 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 texts composed of a single homograph, 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 text. In some cases, more raters were available and we included those in the average (112 cases).

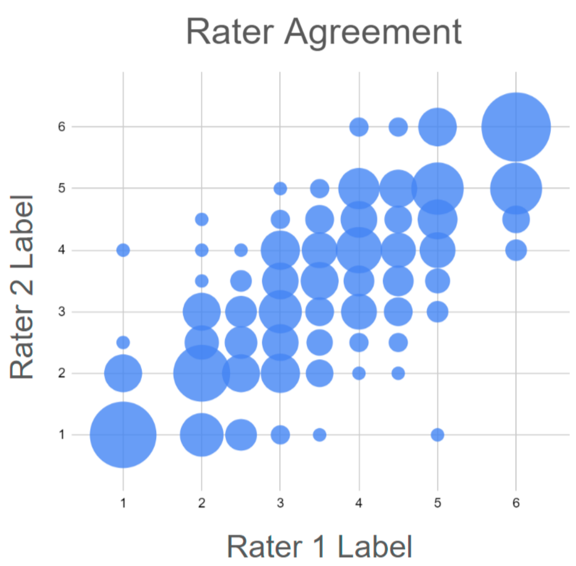

Figure 3: Label agreement between the two primary expert raters. Circle sizes represent the number of texts with each pair of labels. Significantly more disagreement occurs toward the middle of the CEFR scale than at each end.

While most human expert labels were within 1 point of one another, 8% of the labels were further apart than this. Disagreements were particular common for intermediate CEFR levels. Rater agreement is shown in Figure 3. The quadratic weighted kappa (QWK) between the two primary raters is 0.89, which indicates close agreement.

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 for each of the Linear, BERT-based and PaLM 2-L models, the worst 20 predictions from each were re-adjudicated to identify potential mislabels. Results presented in the Experiment section (section 4) are on the final dataset, after all adjudication was completed (123 cases of adjudication in total).

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 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.

For a reference point, we 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.

We took the independent expert labeler MSE of 0.75 (section 4.2) as the main target for machine learning models, although ultimately we were able to surpass the biased metrics of the primary raters as well.

4 Experiment

4.1 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) (Devlin et al., 2018), and a Large Language Model (PaLM

2-L) (Anil et al., 2023) in a few-shot setting. Fine-tuning an LLM was not a focus of this research due to its limited accessibility to many developers, but is a topic of interest for future investigation. As a comparison baseline, the test set was also rated by a human expert. Summary of results is in Figure 4.

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 1.

Table 1: Latency summary of single lookup latency averaged over 100 requests. Latency is estimated within an order of magnitude, and no effort has been made to optimize code for speed. CPU latency was measured on a Linux desktop Intel(R) Xeon(R) CPU E5-2690 v4 @ 2.60GHz with 128 Gb RAM. TPU latency was measured via the Vertex API on a low-latency network connection, querying TPU v5e accelerators. Note that TPU execution is highly parallelizable, so amortized batch lookup speed is substantially faster than individual lookup.

Model Type	Method	Latency (One lookup)
Linear Model on Surface Features	On-device (CPU)	$\sim$50μs
BERT-based Model	On-device (CPU)	$\sim$100ms
BERT-based Model	Via API	$\sim$10ms
PaLM 2-L	Via API	$\sim$1s

4.2 Human Expert

As a basis for comparison, a set of ratings was performed on the test set by a human expert with the same qualifications as the original raters. This expert did not previously work with the labelers of the dataset, but used the rating guideline as well as the training set labels for calibration. Their labels had a MSE of 0.75 (90% confidence [0.67, 0.84]) (Figure 4 (a)).

4.3 Linear Regression Model

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.

4.3.1 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).

The key weakness of these features is that they are content agnostic. For example, "The cat is here." (A1 difficulty) and "His ire is epic." (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 examples in the conversational use case.

4.3.2 Results

Of the models tested, the linear model performed the worst, with an MSE of 0.81 (90% confidence [0.71-0.91]) (Figure 4 (b)). Typical errors relate to mistaking the difficulty of a short word and sentences comprised of short words (Table 3). 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).

4.4 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 text and get a reasonable response. The downside is that these models are comparatively slow (Table 1) 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 for future research.

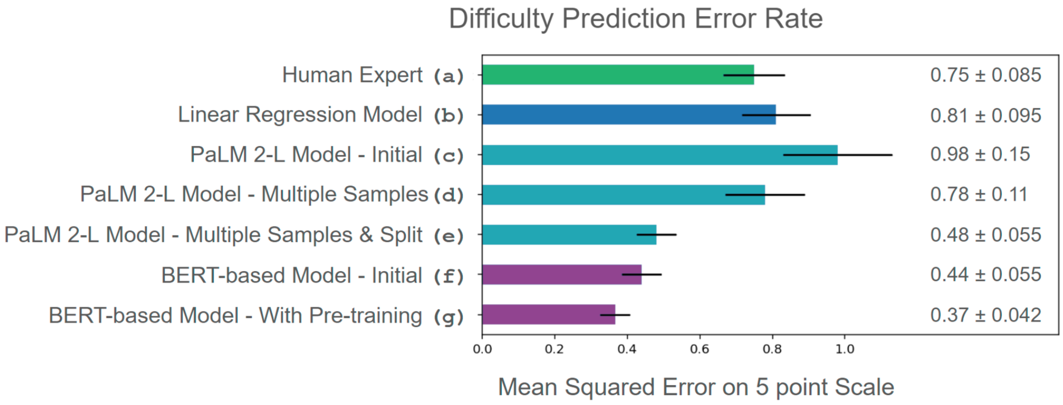

Figure 4: Summary of mean squared error for different model types and training iterations, with 90% confidence intervals. See Section 4 for detailed results and analysis.

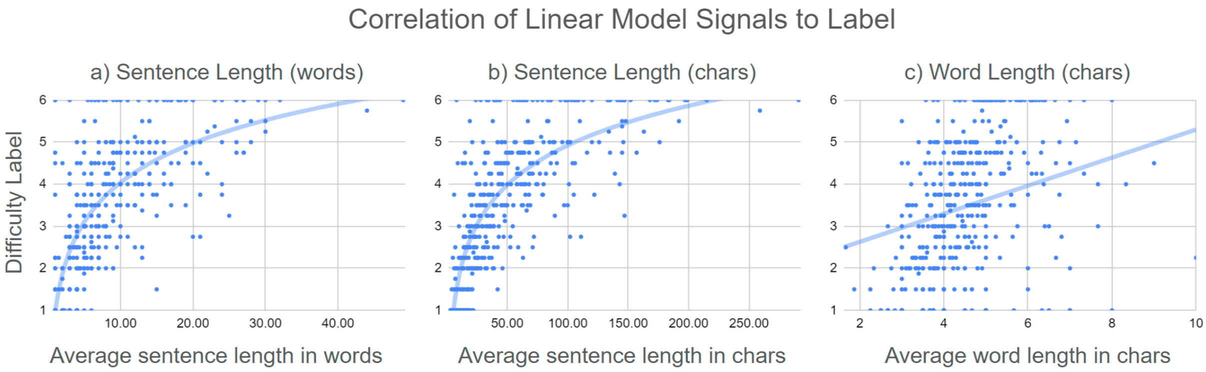

Figure 5: Correlation between linear model signals and label on train set. Correlations are 0.67, 0.70, and 0.35 for average sentence length in words, average sentence length in chars, and average word length in chars, respectively. The sentence length signals have a logarithmic relationship to the label,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

4.4.1 Results

For the initial results, we used a single prompt (A), 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 (Figure 4 (c)).

Since the limitation of the context length prevented us from using all of the training data as few-shot examples, we experimented with running the model multiple times, re-sampling the training data for few-shot examples, and averaging the results. By rerunning the model 3 times, we improved accuracy, from an MSE of 0.98 to 0.78 (Figure 4 (d)). Naturally, this results in proportionately increased latency. Further improvement is likely possible if more samples are taken.

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 (Figure 4 (e)).

The final results are an MSE of 0.48 (90% confidence [0.43, 0.54]) (Figure 4 (e)). This 0.33 better than the linear model and 0.27 better than human expert ratings, albeit at a significant latency cost (Table 1). Unlike the linear model, there is no obvious pattern of errors (Table 4). The opacity of mistakes is a risk factor, since this can make it challenging to improve the model further.

4.5 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 the pre-trained BERT-base-uncased checkpoint 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 easily fine-tuned to very specific tasks. It is also well-suited to learn from a larger teacher model, which was used during a quality iteration.

4.5.1 Results

We finetuned 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 and batch size. The best setup retained the first 3 layers, training them with a learning rate of $6e-5$ at batch size 32 for 6 epochs. The final model has 45.7M parameters and achieved an MSE of about 0.44 (Figure 4 (f)), which is substantially better than any of the other models.

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 10,000 examples from various sources with our best LLM version. We used those LLM-labeled examples to finetune the BERT model using a smaller learning rate of $2e-5$. 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 (Figure 4 (g)).

The final results are an MSE of 0.37 (90% confidence [0.32, 0.41)] (Figure 4 (g)), which is a 0.38 better than the human expert. The latency, particularly when running on TPU (Table 1), 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 Ace-CEFR dataset, but they arose naturally from several of our sources. Ultimately, we decided to correct the misspellings, because we want the dataset to be usable for generative tuning, and mistakes in the input could cause an LLM to learn 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 5)

4.6 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 semantics, 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.

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 pretraining datasets for improvements to BERT in the future.

5 Conclusion

Ultimately, we were able to achieve accuracy better than expert human ratings on short conversational pieces of text. We are releasing the Ace-CEFR 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.

6 Limitations

The Ace-CEFR dataset provides the ability to train models on conversational 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. Particularly impactful is the L1 of the learner, which greatly affects both overall learning difficulty and specific skill acquisition (Ellis, 1985). For example, because French and English have many more cognates than Arabic and English, an L1 speaker of French will likely find different areas of challenge when learning English than an L1 speaker of Arabic. This makes a single scale of difficulty for the two learners to be imperfect for either learner. 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.

A broader inequity inherent to automated tools is the unequal availability of technology to learners of different demographics. Access to computers or mobile phones is not available to everyone, and the demographics that have the most difficulty getting traditional second language education are also likely the ones who will have the least access to computers and mobile phones capable of accessing LLM-based applications for learning. It is important to consider how to maximize accessibility when building applications on top of these technologies, for example, by making them compatible with entry-level consumer devices.

7 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. Finetuning the PaLM 2-L can also be insightful to compare the results against few-shot prompting. Other improvements could be using an LLM with a longer context to include more examples, and cross-training with other datasets such as CEFR-SP. 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

Global scale of english. Online.

Desislava Aleksandrova and Vincent Pouliot. 2023. CEFR-based contextual lexical complexity classifier in English and French. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 518–527, Toronto, Canada. Association for Computational Linguistics.

Bharat Ram Ambati, Siva Reddy, and Mark Steedman. 2016. Assessing relative sentence complexity using an incremental ccg parser. In *15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1051–1057. Association for Computational Linguistics.

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.

Yuki Arase, Satoru Uchida, and Tomoyuki Kajiwara. 2022. Cefr-based sentence difficulty annotation and assessment. *arXiv preprint arXiv:2210.11766*.

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

Ion Madrazo Azpiazu and Maria Soledad Pera. 2019. Multiattentive recurrent neural network architecture for multilingual readability assessment. *Transactions of the Association for Computational Linguistics*, 7:421–436.

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).

Dominique Brunato, Lorenzo De Mattei, Felice Dell'Orletta, Benedetta Iavarone, Giulia Venturi, et al. 2018. Is this sentence difficult? do you agree? In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, pages 2690–2699. Association for Computational Linguistics.

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.

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

Jeanne Sternlicht Chall and Edgar Dale. 1995. Readability revisited: The new dale-chall readability formula. *(No Title)*.

Kiyomi Chujo, Kathryn Oghigian, and Shiro Akasegawa. 2015. A corpus and grammatical browsing system for remedial efl learners. *Multiple affordances of language corpora for data-driven learning*, pages 109–128.

Kevyn Collins-Thompson and James P Callan. 2004. A language modeling approach to predicting reading difficulty. In *Proceedings of the human language technology conference of the North American chapter of the association for computational linguistics: HLT-NAACL 2004*, pages 193–200.

Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2011. Read–it: Assessing readability of italian texts with a view to text simplification. In *Proceedings of the second workshop on speech and language processing for assistive technologies*, pages 73–83.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

R. Ellis. 1985. Understanding second language acquisition.

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.

Edward Fry. 1990. A readability formula for short passages. *Journal of Reading*, 33(8):594–597.

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

Joseph Marvin Imperial. 2021. Bert embeddings for automatic readability assessment. *arXiv preprint arXiv:2106.07935*.

Yasutake Ishii and Yukio Tono. 2018. Investigating japanese efl learners' overuse/underuse of english grammar categories and their relevance to cefr levels. In *Proceedings of the 4th Asia Pacific Corpus Linguistics Conference*, pages 160–165.

Chao Jiang, Mounica Maddela, Wuwei Lan, Yang Zhong, and Wei Xu. 2020. Neural crf model for sentence alignment in text simplification. *arXiv preprint arXiv:2005.02324*.

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.

Matej Martinc, Senja Pollak, and Marko Robnik-Šikonja. 2021. Supervised and unsupervised neural approaches to text readability. *Computational Linguistics*, 47(1):141–179.

Emilio Matricciani. 2023. Readability indices do not say it all on a text readability. *Analytics*.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, et al. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97.

Changping Meng, Muhao Chen, Jie Mao, and Jennifer Neville. 2020. Readnet: A hierarchical transformer framework for web article readability analysis. In *Advances in Information Retrieval: 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14–17, 2020, Proceedings, Part I 42*, pages 33–49. Springer.

Adam Montgomerie. 2021. Attempting to predict the cefr level of english texts. Online.

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

Musa Nushi and Mohammad Hadi Fadaei. 2020. Newsela: A level-adaptive app to improve reading ability.

Sarah E Petersen and Mari Ostendorf. 2009. A machine learning approach to reading level assessment. *Computer speech & language*, 23(1):89–106.

Taraka Rama and Sowmya Vajjala. 2021. Are pre-trained text representations useful for multilingual and multi-dimensional language proficiency modeling? *arXiv preprint arXiv:2102.12971*.

Angeliki Salamoura and Nick Saville. 2010. Exemplifying the cefr: Criterial features of written learner english from the english profile programme. *Communicative proficiency and linguistic development: Intersections between SLA and language testing research*, 1:101–132.

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.

Sanja Štajner, Simone Paolo Ponzetto, and Heiner Stuckenschmidt. 2017. Automatic assessment of absolute sentence complexity. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence, IJCAI*, volume 17, pages 4096–4102.

AJ Stenner, Ivan Horabin, Dean R Smith, and Malbert Smith. 1988. The lexile framework. durham, nc: Metametrics.

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.

Satoru Uchida and Masashi Negishi. 2018. Assigning cefr-j levels to english texts based on textual features. In *Proceedings of Asia Pacific Corpus Linguistics Conference*, volume 4, pages 463–467.

Sowmya Vajjala and Ivana Lučić. 2018. Onestopenglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications*, pages 297–304.

Sowmya Vajjala and Detmar Meurers. 2014. Assessing the relative reading level of sentence pairs for text simplification. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 288–297.

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 (also initially used for words)

```
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 2: **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 3: **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 4: **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 5: **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