Jingle BERT, Frozen All the Way: Freezing Layers to Identify CEFR Levels of Second Language Learners Using BERT

Ricardo Muñoz Sánchez, David Alfter, Simon Dobnik, Maria Irena Szawerna, Elena Volodina
University of Gothenburg, Sweden
ricardo.munoz.sanchez@gu.se

Abstract

In this paper, we investigate the question of how much domain adaptation is needed for the task of automatic essay assessment by freezing layers in BERT models. We test our methodology on three different graded language corpora (English, French and Swedish) and find that partially fine-tuning base models improves performance over fully fine-tuning base models, although the number of layers to freeze differs by language. We also look at the effect of freezing layers on different grades in the corpora and find that different layers are important for different grade levels. Finally, our results represent a new state-of-the-art in automatic essay classification for the three languages under investigation.

1 Introduction

Automated essay scoring (AES) is the “process of evaluating and scoring written prose via computer programs" (Shermis and Burstein, 2003). Even though the implied use of computers nowadays might suggest so, AES is not a recent phenomenon. Ellis Batten Page, also known as “the father of AES" (Wresch, 1993), started to develop his ideas in the 60’s (Page, 1966; Page and Paulus, 1968) and implemented a rather sophisticated program to analyze and grade student essays. Even though work on AES started around 55 years ago, it is still an active area of research to this day (e.g. Beigman Klebanov and Madnani, 2020; Wilkens et al., 2023; Lagutina et al., 2023).

When dealing with pretrained language models, two of the most common approaches are to fine-tune the whole model or to just train any extra classification layers that have been added. Despite that, there have been studies that show that partly fine-tuning the models allows for better domain adaptation by maintaining part of the original knowledge of the model while learning domain-specific features at the same time (Zhu et al., 2021). The reason for this is that different layers of neural models encode different kinds of features, with the first few encoding lower-level features and the later ones encoding higher-level features.

In this paper we aim to determine how much domain adaptation is required for AES. We limit our experiments to BERT models for a couple of reasons. There has been a lot of studies focusing on which layers of these models encode which aspects of linguistic knowledge (e.g. Clark et al., 2019; Jawahar et al., 2019). On the other hand, the more recent generative decoder-only models tend to vary a lot from each other, which can complicate both comparison among themselves and between different languages. Finally, the performance of these decoder-only models in terms of second-language assessment has had mixed results so far (Naismith et al., 2023; Yancey et al., 2023), which in turn means that BERT-based models are still an important part of AES for second language assessment.

Thus we analyze which layers of a pretrained BERT model are important for the task at hand and which ones should be fine-tuned. We assume that the knowledge embedded in the frozen layers (semantics, syntax, grammaticality, etc.) is important for the model to properly determine the proficiency level an essay has been annotated as. We further analyze whether this varies depending on the CEFR level of the essays. That is, we want to determine whether the same encoded knowledge of the language model is equally important for all levels.

We work with the CEFR[1] framework (COE, 2001). It is used to evaluate foreign/second language learning by assigning one of the six levels (A1, A2, B1, B2, C1, C2) that determine the proficiency of second language (L2) speakers. Furthermore, we work with three different languages: English, French and Swedish. While CEFR-labeled

[1] Common European Framework of Reference for Languages

Ricardo Muñoz Sánchez, David Alfter, Simon Dobnik, Maria Szawerna and Elena Volodina. Jingle BERT, Frozen All the Way: Freezing Layers to Identify CEFR Levels of Second Language Learners Using BERT. *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning (NLP4CALL 2024).*
Linköping Electronic Conference Proceedings 211: 137–152.
137

data can be scarce, there is a growing societal need for automated grading in CEFR terms. An example of this is how different governments are either planning to require a language test for applicants for residence and citizenship or already do so (Code civil français, 2011; Swedish Government, 2021, 2023; U.S. Citizenship and Immigration Services, 2023; Government of Canada, 2024). Because of this, we expect that the need for support in AES will drastically increase in the near future, both as a way to support self-studying learners and for high-stakes essay grading.

The rest of our paper is organized as follows. Section 2 introduces the context for our experiment in terms of previous research. In Section 3.2 we describe our approach, as well as the considerations we have taken into account while designing it. Section 3.1 describes the datasets used for our experiments, while Section 3.3 describes the state-of-the-art we compare our models to. We present our results as well as a discussion of these in Section 4. Finally, we present our conclusions in Section 5, as well as possible directions in which to expand our work.

2 Related Work

The state-of-the-art in AES has long been dominated by systems using feature engineering and linguistic variables that measure textual quality, such as number of words (Shermis and Burstein, 2003; Parslow, 2015), number of grammatical errors (Yannakoudakis et al., 2018; Ballier et al., 2019), type-token ratio (Vajjala and Lõo, 2014; Lee and Hasebe, 2020), or lexical density (Hancke, 2013; Hancke and Meurers, 2013; Pilán and Volodina, 2018). It is only recently that deep learning approaches have begun to set new standards (Hussein et al., 2019; Bestgen, 2020).

Alikaniotis et al. (2016) and Taghipour and Ng (2016) were the first ones to use deep learning for AES. Even though they used an LSTM[2] architecture (Hochreiter and Schmidhuber, 1997), other network architectures such as Convolutional Neural Networks (CNN) and Recurrent Convolutional Neural Networks (RCNN) have also been successfully applied in the past (Dong and Zhang, 2016; Dong et al., 2017; Dasgupta et al., 2018; Shin and Gierl, 2021).

Recent experiments using GPT for CEFR classification have found that GPT-4 (OpenAI, 2024) can reach performances approaching those of sophisticated automated scoring systems (Banno et al., 2024), although agreement with human annotators remained inconclusive (Yancey et al., 2023). Large Language Models have also been used for other tasks related to computational approaches to language learning, such as learner-adapted definition generation (Yuan et al., 2022), learner-centered text simplification (Baez and Saggion, 2023), or proficiency-adapted text generation (Bezirhan and von Davier, 2023).

As with most fields in NLP, most of the work in this field has been done in English (Søgaard, 2022). A consequence of that is that other languages are often not paid enough attention to.

For instance, very little work has been done on essay classification in Swedish, some examples being Östling et al. (2013) on grading upper-secondary essays written by native speakers, Pilán (2018) on CEFR classification of L2 learner essays, Lilja (2018) on assigning grades to high-school essays, and Ruan (2020) on assigning grades to essays written as a part of national exams. Some of these works use the Uppsala Corpus of Student Writings (Megyesi et al., 2016). This corpus mainly consists of native speaker upper secondary level writings but also contains some texts, around 8%, written by learners of Swedish as a second language. However, it is not aligned with the CEFR scale.

Both Lilja (2018) and Ruan (2020) use deep learning to classify these essays by assigned grades. Lilja (2018) uses an LSTM and explores whether pre-trained embeddings are better or not than a fine-tuned version or randomly initialized ones. They conclude that pre-trained fine-tuned embeddings produce the best results, but due to high standard deviations, they are not significantly different from randomly initialized embeddings.

Ruan (2020), explores the use of hand-crafted features in combination with deep neural networks. The feature categories are virtually identical to those in Pilán (2018), namely count-based, morphological, syntactic and lexical. Semantic features were not included. The chosen architecture is a recurrent neural network. Using each feature group separately, they find that all feature groups perform similarly, although each feature group separately performs better than using all features simultaneously. Overall, they find that a feature-based system outperforms the word embeddings based system by Lilja (2018).

[2] Long Short-Term Memory

A similar situation presents itself for French, with a limited number of studies on essay classification. For non-L2 French, Lemaire and Dessus (2001) use Latent Semantic Analysis to grade a limited number of student essays (31), and Zaghouani (2002) presents a conceptual design for grading essays using a multi-agent system. Parslow (2015) presents a preliminary study on automatic grading of L2 French essays written by Swedish native speakers using feature-based methods and Naive Bayes classifiers. Finally, Ranković et al. (2020) use CamemBERT to extract word-level features and a deep recurrent network to grade essays written by French learners in German-speaking parts of Switzerland.

Mayfield and Black (2020) argue that the move to deep neural models for AES comes with considerable computational costs while producing performance comparable to the classical models. Their conclusions indicate, however, that there is a further need to explore deep learning approaches.

3 Materials and Methods

Second language assessment is a high-stakes situation, given that its outcome can affect the educational and professional opportunities that a student has available to them. While deep learning models tend to out-perform feature-based models, they tend to be obscure, with little to no explanation both of where specific predictions come from and which kind of features they focus on (Guidotti et al., 2018).

In this section, we first introduce the datasets we used in Section 3.1, followed by our approach to obtain a more explainable BERT (Devlin et al., 2019) model in Section 3.2. Finally, we talk about the state-of-the-art we compare our approach to in Section 3.3.

3.1 Datasets

3.1.1 English Dataset

We are using the EFCamDat corpus (Geertzen et al., 2013) for experiments on English. The corpus consists of essays collected from the EF Education First online platform. The essays were assigned a grade on a 16-level scale with equivalents to some of the major standards in L2 language learning, including CEFR levels. However, it should be noted that the grades were assigned according to the level the students reached in the platform as opposed to direct evaluation of the essays themselves.

Level	# essays	# train	# valid	# test
A1	192K	2,299	767	767
A2	130K	1,555	518	518
B1	62K	738	246	246
B2	18K	218	73	73
C1	5K	62	20	20
C2	0	0	0	0
Total	406K	4,872	1,624	1,624

Table 1: Number of essays in the English L2 learner corpus (EFCamDat) for each of the CEFR levels. The letter *K* denotes that the numbers we are dealing are in the thousands. Note that there are no C2 level essays in the corpus. We randomly sample a small percentage of the corpus for faster training while keeping the label distributions the same.

The corpus contains over 400,000 essays from CEFR levels ranging from A1 to C1, as seen in Table 1. The students are placed into one of the platform's 16 levels either through a placement test or by progressing through the course. Each level has eight possible writing tasks, which gives a wide array of possible topics for each CEFR level. Given that we are training the models several times, we sampled 2% of the data to keep the use of computational resources within a reasonable margin. The essays were randomly sampled and stratified by CEFR level, to maintain the proportion of each label. Moreover, this leaves us with a dataset of a comparable size to TCFLE-8, the French corpus we are using.

3.1.2 French Dataset

For French, we use the recently released TCFLE-8 corpus (Wilkens et al., 2023). This is a corpus based on the French language certification exam TCF (test de connaissance du français 'French knowledge test') administered by the France Éducation International. It is the biggest French corpus for AES to date with over 6.5k essays and covers a wide variety of prompts.

All essays are graded by at least 2 professional raters and cover all six levels of the CEFR scale, as seen in Table 2. Different data cleaning and quality assurance steps were taken by the corpus creators to ensure that the corpus contains representative samples at each level.

Level	# essays	# train	# valid	# test
A1	689	413	138	138
A2	1,375	825	275	275
B1	1,466	880	293	293
B2	1,427	856	285	285
C1	1,127	676	226	226
C2	485	0	0	0
Total	6,569	3,650	1,217	1,217

Table 2: Number of essays in the French L2 learner corpus (TCFLE-8) for each of the CEFR levels. Note that this is the only corpus of the three that we are working with that contains C2 level essays. We have removed the essays of this level to allow for better comparison across languages.

Level	# essays	# train	# valid	# test
A1	59	35	12	12
A2	143	85	29	29
B1	86	52	17	17
B2	105	63	21	21
C1	96	58	19	19
C2	7	0	0	0
Missing	6	0	0	0
Total	502	293	98	98

Table 3: Number of essays in the Swedish L2 learner corpus (Swell-Pilot) for each of the CEFR levels. Note that there are very few essays of level C2 in the corpus and that some are missing a level.

3.1.3 Swedish Dataset

For Swedish, we use the Swell-pilot corpus (Volodina et al., 2016a; Volodina, 2024). It consists of three subcorpora of L2 Swedish learners (see below) and is annotated with CEFR levels. All CEFR levels are well represented in the corpus, with the exception of C2 level (advanced) essays, as seen in Table 3. Thus we remove the C2 essays as their low number would not be representative of the model's classification capabilities. Moreover, there are six essays that lack a level which have been ignored for the purposes of this experiment.

SpIn consists of 256 essays from a course for refugees that had recently arrived to Sweden. The course was introductory in nature and the essays were part of a mid-term exam.

SW1203 consists of 141 essays from a preparatory course for foreign students that intended to study an undergraduate program in Sweden.

TISUS consists of 105 essays from the written part of the Test In Swedish for University Studies (TISUS)[3]. The essays are argumentative, the topic being “stress".

3.2 Methodology

In order to classify the essays, we use language-specific versions of BERT. For the experiments themselves, we explore how freezing different layers of BERT during training affects its performance. We freeze the layers in a bottom-up manner, given that lower layers learn more basic linguistic features such as surface-level features, while higher layers learn more task-specific features, such as semantic and contextual features (Clark et al., 2019; Jawahar et al., 2019). Thus, we compare different configurations ranging from a completely fine-tuned model to one where only the classification layer was trained.

For the classification task itself, we truncate the essays to fit the maximum token length of BERT and feed them to the model.[4] We then take the top layer representation of the `[CLS]` token and feed it to a linear layer for classification. Taking the output of the same layer all the time allows us to compare the differences between how the models are learning depending on how many layers we have frozen.

In terms of hyperparameters, we explore using different learning rates[5] and find that the best performing on average is 5e-5. We also run the experiments for 10 epochs, loading the best performing checkpoint at the end.

Given that none of the corpora used has standard train/test splits, we run our experiments five times, generating new train/validation/test splits with a 60/20/20 distribution each run to account for variance. We maintain the proportions of the different CEFR levels across the splits. The number of each label per level can be seen in Tables 1, 2, and 3.

As for our models, we use specific versions of BERT according to the language.

For English, we use the original version of BERT[6] (Devlin et al., 2019). It was trained using BooksCorpus (Zhu et al., 2015) and an English Wikipedia dump. Note that we are using the cased

[3] https://www.su.se/tisus/english/

[4] Note that we are not using Longformer as it is not available in all of the languages we are working with.

[5] We experimented with learning rates of 1e-4, 5e-4, 1e-5, 5e-5, 1e-6, 5e-6, and 1e-7.

[6] https://huggingface.co/google-bert/bert-base-cased

version of BERT as the Swedish model has no uncased version available.

We use CamemBERT[7] (Martin et al., 2020) for French, which is based on RoBERTa (Liu et al., 2019) rather than on vanilla BERT. It was trained using the French section of the OSCAR corpus (Suárez et al., 2019), a language annotated version of CommonCrawl.[8]

The final model we use is Swedish BERT[9] (Malmsten et al., 2020), a Swedish version of BERT implemented by KBLab at theNational Library of Sweden[10]. It was trained on a combination of corpora containing newspapers, social media, official reports from the Swedish government, legal documents, and Wikipedia in Swedish.

3.3 State-of-the-art

In this section we talk about the current state-of-the-art in AES within the context of the datasets we are using. These results are summarized in the top row of Table 4.

3.3.1 English

The most similar work to our own for English is by Schmalz and Brutti (2021) who use BERT for the classification of the EFCamDat data. They also work on subsets of the whole data (10k, 50k, 100k) due to space and computational constraints with using the whole corpus. We report their best results as state-of-the-art.[11]

3.3.2 French

Wilkens et al. (2023) perform a series of essay classification experiments on the TCFLE-8 corpus in order to establish some first baselines: (1) a transformer-based approach using CamemBERT, (2) a feature-based approach using XGBoost, and (3) a simple logistic regression. For the feature-based algorithm in (2) and (3) they use a set of 119 features – distilled from over 5k features – from nine subcategories: errors, graded lexicons, lexical diversity, lexical frequency, lexical sophistication, orthographic neighbors, morphology, tenses, likelihood, and word length. They find the transformer-based model to perform best, followed by XGBoost. For brevity, we will only report results from their best-performing model (i.e., the transformer-based model) as state-of-the-art.

3.3.3 Swedish

For Swedish, we compare our model with a feature-based approach to be able to draw a comparison between performance and explainability. Pilán et al. (2016) and Volodina et al. (2016b) use a feature set of about 60 features divided into five subcategories: length-based, lexical, morphological, syntactic, and semantic features. They use an SVM to classify the data. Both studies found that lexical features perform the best.

Pilán and Volodina (2018) specifically investigate the importance of features for the classification of (1) sentences, (2) reading texts from textbooks, and (3) learner essays from SweLL-pilot. Using analysis of variance (ANOVA), they determine the most predictive features for each of the three subgenres of text. In general, this study corroborates findings from Crossley and McNamara (2011) for L2 English in that lexical diversity and lexical frequency are strong predictors in both studies, and Vajjala and Lõo (2014) who also found verb variation and lexical variation to be strong predictors for L2 Estonian.

3.4 Evaluation

We evaluate our system both in terms of accuracy and of “adjacent accuracy". The idea behind adjacent accuracy is that an A1 essay misclassified as A2 is a smaller mistake as opposed to it being misclassified as a B2 essay.

In more formal terms, we say that a prediction is correct in terms of adjacent accuracy if: (1) our classes are ordinal and (2) the prediction is either the correct class or the immediate predecessor or successor of it.

Moreover, we use F1 score calculated using both usual and adjacent accuracy. We report both macro and weighted F1 scores as they aggregate the F1 scores for the individual classes assuming either that the classes are equally important (for macro averaging) or that the number of examples for each class matter (for weighted averaging).

4 Results and Discussion

4.1 Performance Across Languages

In this section we present the results of our experiments, noting the performance across languages

[7] https://huggingface.co/almanach/camembert-base

[8] https://commoncrawl.org/about/

[9] https://huggingface.co/KB/bert-base-swedish-cased

[10] https://www.kb.se/in-english/research-collaboration/kblab.html

[11] It would arguably be fairer to compare against the results they obtained with the smallest subsample, approaching our own sample size.

Layers Frozen	English	French	Swedish
State-of-the-art	0.974	0.56	0.23
None	0.975 ± 0.000	0.555 ± 0.003	0.722 ± 0.018
All layers	0.319 ± 0.000	0.443 ± 0.005	0.188 ± 0.001
Embedding Layer	0.971 ± 0.000	0.526 ± 0.005	0.727 ± 0.008
1 Encoder Layer	0.974 ± 0.000	0.517 ± 0.011	0.731 ± 0.019
1 and 2	0.974 ± 0.000	0.524 ± 0.010	**0.744** ± **0.011**
1 to 3	0.974 ± 0.000	0.538 ± 0.002	0.718 ± 0.006
1 to 4	**0.977** ± **0.000**	0.529 ± 0.011	0.720 ± 0.003
1 to 5	0.972 ± 0.000	0.537 ± 0.008	0.725 ± 0.010
1 to 6	0.966 ± 0.000	0.532 ± 0.017	0.705 ± 0.006
1 to 7	0.967 ± 0.000	0.542 ± 0.018	0.671 ± 0.009
1 to 8	0.962 ± 0.000	0.548 ± 0.006	0.664 ± 0.020
1 to 9	0.957 ± 0.000	0.552 ± 0.004	0.612 ± 0.011
1 to 10	0.946 ± 0.000	0.564 ± 0.004	0.596 ± 0.013
1 to 11	0.919 ± 0.000	**0.572** ± **0.001**	0.541 ± 0.004

Table 4: Weighted F1 scores for the different languages. Even though the number of layers to freeze to obtain the best-performing model varies across languages, the best model is always partially fine-tuned.

and CEFR levels. More detailed tables and results for each language can be found in Appendix B for the metrics based on accuracy and in Appendix C for those based on adjacent accuracy. Table 4 compares the weighted F1 scores among languages.

First of all we can notice that all BERT models that were even partially fine-tuned performed better than the fully frozen model. That is, fine-tuning even one layer led to large improvements in the performance.

Even though the best performing model was always partially fine-tuned, which layers should be frozen varied depending on the language. For instance, for English, the only model that performed better than the fully fine-tuned one was the one where we froze all layers up to the fourth encoder layer, indicating a reliance on surface-level features for classification. Meanwhile, the French model showed a preference towards fine-tuning just the last few encoder layers, indicating that a broad range of linguistic features may be necessary to accurately classify the essays. Finally, the Swedish model worked the best when few of the encoder layers were frozen, which again points to the importance of surface-level features for AES in Swedish.

Based on this, we can assume that maintaining basic knowledge of the language within the model is an important part of automated essay grading. This sounds reasonable, given that second language learners tend to demonstrate an imperfect usage of the language. Moreover, we would prefer not to have this usage of the language overwrite the knowledge of the model.

Something notable is that when the model misclassified an essay, it usually assigned that essay to one of the adjacent levels. Even though the CEFR levels are ordinal to us humans, this information was not provided to the model at any point during training. This points to the model learning how to identify the level of the essay according to the linguistic characteristics, as students from adjacent levels are more likely to create similar texts than those for levels that are farther apart.

4.2 Performance Across CEFR Levels

Figures 1, 2, and 3 show how the different levels react to fine-tuning different layers of the models. We have cut-off the values that are below a certain threshold for each of the plots as they do not help us identify which layers are important for that specific class. Nevertheless, the full figures can be found in Appendix A.

For French and for Swedish we notice that the levels where the model performs the best are those that are closer to the edges of the CEFR scale, regardless of the language. This points to these levels being easier to classify as they are the most likely to be different from the other essays. On the other hand, levels B1 and B2 are the ones that have lower

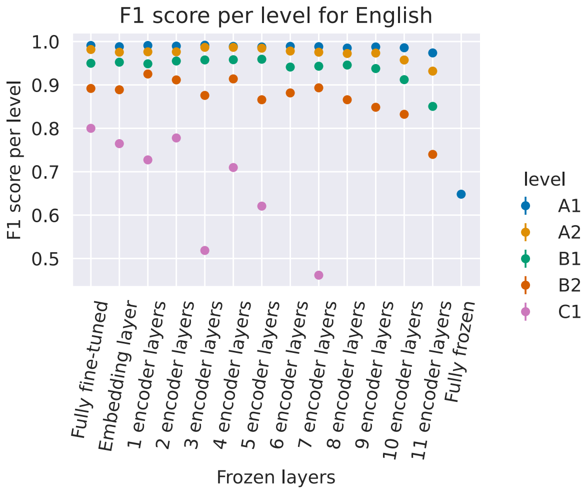

Figure 1: Performance per CEFR level when freezing different layers of BERT. Note that the performance tends to drop as the levels increase.

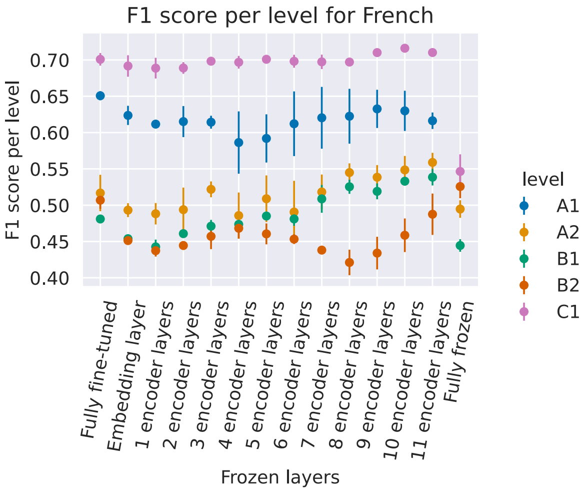

Figure 2: Performance per CEFR level when freezing different layers of CamemBERT. Note that even though all levels perform differently, most of them have a slight uptick in performance when we finetune only the last few encoder layers.

F1 scores. This might be due to them being more similar to their adjacent levels and thus harder to properly identify.

In more language specific notes, we see that the different levels tend to follow the same trend as the overall performance of each model.

We begin by looking at how the English BERT behaves across levels in Figure 1. We can note that the performance is inversely correlated to the level. That is, lower levels get higher F1 scores, while higher levels get lower F1 scores. This might be due to the prompts given to the students. For example, A1 essays have an almost perfect classification. However, most of them begin with a salutation (hi, hello, etc.) and address someone called Anna. This could in turn lead to leakage, which would explain the high performance seen in Table 4 compared to French and Swedish. Moreover, the levels are inferred from the course level, which Muñoz Sánchez et al. (2024b) argue is not necessarily a good proxy for CEFR levels. As for the individual levels, we notice that the general trend is for their accuracy to drop the more layers we freeze. Even though there are some layers that have either higher or lower perplexity, they do not seem to follow a pattern.

When looking at the French model in Figure 2 we notice that most of the levels have a slight increase in their performance as we approach the latter layers. However, different levels behave differently. For instance, the performance for level C1 is mostly stable with a very slight decrease when freezing just the first few layers and a very slight increase when fine-tuning just the last few layers. Meanwhile, level A1 has its highest performance when fine-tuning all of the model and another increase when freezing layers up to the ninth or tenth encoder layers, which points to the importance of a broad range of features. With levels A2, B1, and B2 we see a similar pattern: fine-tuning the whole model leads to higher performance but fine-tuning just the final encoder layer leads to the highest performance for these levels. Thus, we can assume that low-, mid- and high-level features play an important role in French AES. Even though the performance of our best model is similar to the one reported by Wilkens et al. (2023), we still see an increase in performance when freezing layers compared to fully fine-tuning the base model.

Finally, we take a look at Swedish BERT in Figure 3. Here we notice that there are two humps in the performance for levels A1 and A2. The first is when freezing just the first few layers and the second one is when freezing up to the first four or five encoder layers. This points to the importance of lexical and syntactic features. A similar pattern can be observed for level B1, albeit in a more erratic manner. For levels B2 and C1 we notice that freezing the first two decoder layers leads to the highest performance, pointing to the importance of lexical features.

5 Conclusions and Future Work

In this study we analyzed different fine-tuning strategies for AES using BERT-based models.

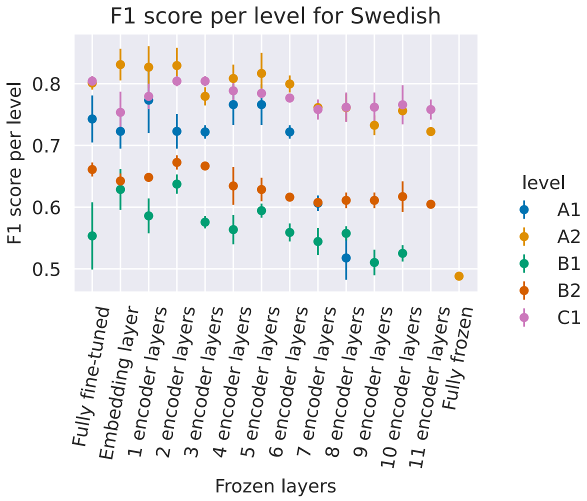

Figure 3: Performance per CEFR level when freezing different layers of Swedish BERT. Note that even though all levels perform differently, most of them have a sharp drop in performance when we finetune only the last few encoder layers.

Even though there was no unified pattern across languages on which layers are crucial, we show that the best-performing modes are ones that have gone through domain adaptation by partial fine-tuning. We also show that even though the importance of layers when taking into account the performance on each individual class differs, it tends to closely follow that of the whole model.

There are several directions in which our work can be expanded upon. The most immediate one would be to expand the languages used, as this would allow to identify if there are patterns depending on language families. On a similar note, we would be interested in seeing the effects of the L1 of a student on which layers and/or features are more important for the assessment.

Another important follow-up of our work would be to determine whether freezing specific layers leads to more fair systems. The idea behind this would be that a fair model should focus on the knowledge and skills of the students as opposed to spurious correlations such as (indirectly) using demographic data for classification. Human graders do tend to show slight biases based on these characteristics (Aldrin, 2017) and study on how deep learning models deal with these has been limited to perceived ethnicity of names (Muñoz Sánchez et al., 2024a).

Finally, we consider that it is important to do a deeper analysis both of the terms appearing in the essays and of the kinds of prompts given to the students. As we mentioned, almost all of the essays in the A1 level in the English dataset include salutations as their first word. This is because the prompts for this level ask the students to greet or to introduce themselves to someone in specific. This can lead to a dataset in which it is not easy to identify whether our model is behaving as we expect or if it is looking as spurious correlations.

We consider that this work is an important step towards understanding which features are important when using transformer-based models for AES. This will in turn help create better and more interpretable models for this task, as well as will contribute to their fairness.

Limitations

The present work only reports on works for the automatic assessment of written language. It should be mentioned that there is a substantial body of work done on automatic assessment of speech as well. Speech has its own specificities, for example fluency. Fluency is the rate at which one speaks, as operationalized in the Complexity, Accuracy, Fluency (CAF) framework (Skehan et al., 1998).

On top of that, the datasets and the approach we use in this paper aggregate several characteristics such as the grammar, vocabulary, relevance, among others into a single label for the whole essay. Naismith et al. (2023) note that this can lead to issues when automatically assigning a level to the essay, as some of these characteristics are harder to capture computationally, such as discourse coherence.

Another thing to note is that the models we used were originally trained using vastly different amounts of data. This could lead to differences in how they model language. For example, the models performed extremely well for the English dataset, while the performance was lower for both the French and the Swedish datasets. We recommend further analysis and cross-examination to ensure that none of these datasets were included in the training data for any of these models. On top of that, the French model is based on RoBERTa not on BERT, which might affect the results. To the best of out knowledge, CamemBERT is the most commonly used model derived from BERT in French.

Ethics Statement

It is important to note that our model should not be used as a substitute for expert human graders.

As noted during the results, not even our model achieves perfect accuracy, which could impact the lives of students. Thus, we suggest always keeping a human-in-the-loop approach with this kind of technology.

Acknowledgements

The research presented here has been enabled by the Swedish national research infrastructure Nationella Språkbanken, funded jointly by the Swedish Research Council (2018–2024, contract 2017-00626) and the 10 participating partner institutions.

This work is part of a continuous endeavour on exploring the use of machine learning techniques to study second language learner texts in Swedish. Moreover, we aim to modernise the L2 tools that are currently available through the research infrastructure of Språkbanken Text, in particular via Språkbanken's learning platform Lärka,[12] which allows researchers, as well as teachers and learners, to interact and analyse these kinds of texts in an automated manner.

The third author is also supported by a grant from the Swedish Research Council (VR project 2014-39) for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg.

References

Emilia Aldrin. 2017. Assessing Names? Effects of Name-Based Stereotypes on Teachers' Evaluations of Pupils' Texts. *Names*, 65(1):3–14.

Dimitrios Alikaniotis, Helen Yannakoudakis, and Marek Rei. 2016. Automatic text scoring using neural networks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 715–725, Berlin, Germany. Association for Computational Linguistics.

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.

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 Maren Scheffel, Julien Broisin, Viktoria Pammer-Schindler, Andri Ioannou, and Jan Schneider, editors, *Transforming Learning with Meaningful Technologies*, volume 11722, pages 308–320. Springer International Publishing, Cham. Series Title: Lecture Notes in Computer Science.

Stefano Banno, Hari Krishna Vydana, Kate Knill, and Mark Gales. 2024. Can GPT-4 do L2 analytic assessment? In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 149–164, Mexico City, Mexico. Association for Computational Linguistics.

Beata Beigman Klebanov and Nitin Madnani. 2020. Automated evaluation of writing – 50 years and counting. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7796–7810, Online. Association for Computational Linguistics.

Yves Bestgen. 2020. Reproducing monolingual, multilingual and cross-lingual CEFR predictions. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5595–5602, Marseille, France. European Language Resources Association.

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

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Code civil français. 2011. Article 21-24 (version en vigueur depuis le 18 juin 2011 [entered into force on june 18, 2011]).

Council of Europe. COE. 2001. *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.

Scott A. Crossley and Danielle S. McNamara. 2011. Understanding expert ratings of essay quality: Coh-Metrix analyses of first and second language writing. *International Journal of Continuing Engineering Education and Life Long Learning*, 21(2-3):170–191.

Tirthankar Dasgupta, Abir Naskar, Lipika Dey, and Rupsa Saha. 2018. Augmenting textual qualitative features in deep convolution recurrent neural network for automatic essay scoring. In *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 93–102, Melbourne, Australia. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for*

[12] https://spraakbanken.gu.se/larka/

Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Fei Dong and Yue Zhang. 2016. Automatic features for essay scoring – an empirical study. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1072–1077, Austin, Texas. Association for Computational Linguistics.

Fei Dong, Yue Zhang, and Jie Yang. 2017. Attention-based recurrent convolutional neural network for automatic essay scoring. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 153–162, Vancouver, Canada. Association for Computational Linguistics.

Jeroen Geertzen, Theodora Alexopoulou, Anna Korhonen, et al. 2013. Automatic linguistic annotation of large scale l2 databases: The ef-cambridge open language database (efcamdat). In *Proceedings of the 31st Second Language Research Forum. Somerville, MA: Cascadilla Proceedings Project*, pages 240–254.

Government of Canada. 2024. Documents for Express Entry: Language requirements. Accessed: 14-06-2024.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42.

J Hancke. 2013. Automatic prediction of CEFR proficiency levels based on linguistic features of learner language. *Master Thesis. University of Tübingen, Tübingen, Germany*.

Julia Hancke and Detmar Meurers. 2013. Exploring CEFR classification for German based on rich linguistic modeling. In *Proceedings of the Learner Corpus Research (LCR) conference*, pages 54–56.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-term Memory. *Neural computation*, 9:1735–80.

Mohamed Abdellatif Hussein, Hesham Hassan, and Mohammad Nassef. 2019. Automated language essay scoring systems: a literature review. *PeerJ Computer Science*, 5(208):1–16.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Nadezhda Stanislavovna Lagutina, Kseniya Vladimirovna Lagutina, Anastasya Mikhailovna Brederman, and Natalia Nikolaevna Kasatkina. 2023. Text classification by cefr levels using machine learning methods and bert language model. *Modelirovanie i Analiz Informatsionnykh Sistem*, 30(3):202–213.

Jae-Ho Lee and Yoichiro Hasebe. 2020. Quantitative Analysis of JFL Learners' Writing Abilities and the Development of a Computational System to Estimate Writing Proficiency. *Learner Corpus Studies in Asia and the World*, 5:105–120.

Benoit Lemaire and Philippe Dessus. 2001. A System to Assess the Semantic Content of Student Essays. *Journal of Educational Computing Research*, 24(3):305–320.

Mathias Lilja. 2018. *Automatic Essay Scoring of Swedish Essays using Neural Networks*. PhD Thesis. Uppsala University.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Martin Malmsten, Love Börjeson, and Chris Haffenden. 2020. Playing with words at the national library of sweden – making a swedish bert.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Elijah Mayfield and Alan W Black. 2020. Should you fine-tune BERT for automated essay scoring? In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 151–162, Seattle, WA, USA → Online. Association for Computational Linguistics.

Beáta Megyesi, Jesper Näsman, and Anne Palmér. 2016. The Uppsala corpus of student writings: Corpus creation, annotation, and analysis. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3192–3199, Portorož, Slovenia. European Language Resources Association (ELRA).

Ricardo Muñoz Sánchez, Simon Dobnik, Maria Irena Szawerna, Therese Lindström Tiedemann, and Elena Volodina. 2024a. Did the names I used within my essay affect my score? diagnosing name biases in automated essay scoring. In *Proceedings of the Workshop on Computational Approaches to Language Data Pseudonymization (CALD-pseudo 2024)*, pages 81–91, St. Julian's, Malta. Association for Computational Linguistics.

Ricardo Muñoz Sánchez, Simon Dobnik, and Elena Volodina. 2024b. Harnessing GPT to study second language learner essays: Can we use perplexity to determine linguistic competence? In *Proceedings* *of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 414–427, Mexico City, Mexico. Association for Computational Linguistics.

Ben Naismith, Phoebe Mulcaire, and Jill Burstein. 2023. Automated evaluation of written discourse coherence using GPT-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 394–403, Toronto, Canada. Association for Computational Linguistics.

OpenAI. 2024. Gpt-4 technical report.

Robert Östling, Andre Smolentzov, Björn Tyrefors Hinnerich, and Erik Höglin. 2013. Automated essay scoring for Swedish. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 42–47, Atlanta, Georgia. Association for Computational Linguistics.

Ellis B. Page. 1966. The imminence of... grading essays by computer. *The Phi Delta Kappan*, 47(5):238–243. Publisher: JSTOR.

Ellis B. Page and Dieter H. Paulus. 1968. The Analysis of Essays by Computer. Final Report. Technical report, The University of Connecticut.

Nicholas Parslow. 2015. Automated Analysis of L2 French Writing: a preliminary study. Master's thesis. Publisher: Unpublished.

Ildikó Pilán and Elena Volodina. 2018. Investigating the importance of linguistic complexity features across different datasets related to language learning. In *Proceedings of the Workshop on Linguistic Complexity and Natural Language Processing*, pages 49–58, Santa Fe, New-Mexico. Association for Computational Linguistics.

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, Osaka, Japan. The COLING 2016 Organizing Committee.

Ildikó Pilán. 2018. *Automatic proficiency level prediction for Intelligent Computer-Assisted Language Learning*. PhD Thesis, University of Gothenburg, Gothenburg, Sweden.

Bojana Ranković, Sarah Smirnow, Martin Jaggi, and Martin J. Tomasik. 2020. Automated Essay Scoring in Foreign Language Students Based on Deep Contextualised Word Representations. In *LAK20-10th International Conference on Learning Analytics & Knowledge*. Issue: CONF.

Rex Dajun Ruan. 2020. *Neural Network Based Automatic Essay Scoring for Swedish*. Master Thesis. Uppsala University.

Veronica Juliana Schmalz and Alessio Brutti. 2021. Automatic assessment of English CEFR levels using BERT embeddings. In *Proceedings of the Eighth Italian Conference on Computational Linguistics*.

Mark D. Shermis and Jill C. Burstein. 2003. *Automated essay scoring: A cross-disciplinary perspective*. Routledge.

Jinnie Shin and Mark J. Gierl. 2021. More efficient processes for creating automated essay scoring frameworks: A demonstration of two algorithms. *Language Testing*, 38(2):247–272.

Peter Skehan et al. 1998. *A cognitive approach to language learning*. Oxford University Press.

Anders Søgaard. 2022. Should we ban English NLP for a year? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5254–5260, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. In *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019*, pages 9 – 16, Mannheim. Leibniz-Institut für Deutsche Sprache.

Svenska Regering Swedish Government. 2021. Krav på kunskaper i svenska och samhällskunskap för svenskt medborgarskap, sou 2021:2.

Svenska Regering Swedish Government. 2023. Kunskapskrav för permanent uppehållstillstånd, sou 2023:25.

Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1882–1891, Austin, Texas. Association for Computational Linguistics.

U.S. Citizenship and Immigration Services. 2023. USCIS Policy Manual: Chapter 2 - English and Civics Testing. Accessed: 14-06-2024.

Sowmya Vajjala and Kaidi Lõo. 2014. Automatic CEFR level prediction for Estonian learner text. In *Proceedings of the third workshop on NLP for computer-assisted language learning*, pages 113–127, Uppsala, Sweden. LiU Electronic Press.

Elena Volodina. 2024. On two SweLL learner corpora – SweLL-pilot and SweLL-gold. *Huminfra Conference*, pages 83–94.

Elena Volodina, Ildikó Pilán, Ingegerd Enström, Lorena Llozhi, Peter Lundkvist, Gunlög Sundberg, and Monica Sandell. 2016a. SweLL on the rise: Swedish learner language corpus for European reference level studies. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 206–212, Portorož, Slovenia. European Language Resources Association (ELRA).

Elena Volodina, Ildikó Pilán, and David Alfter. 2016b. Classification of Swedish learner essays by CEFR levels. In *CALL communities and culture – short papers from EUROCALL 2016*, pages 456–461. Research-publishing.net.

Rodrigo Wilkens, Alice Pintard, David Alfter, Vincent Folny, and Thomas François. 2023. TCFLE-8: a corpus of learner written productions for French as a foreign language and its application to automated essay scoring. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3447–3465, Singapore. Association for Computational Linguistics.

William Wresch. 1993. The imminence of grading essays by computer—25 years later. *Computers and Composition*, 10(2):45–58.

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, Toronto, Canada. Association for Computational Linguistics.

Helen Yannakoudakis, Øistein E. Andersen, Ardeshir Geranpayeh, Ted Briscoe, and Diane Nicholls. 2018. Developing an automated writing placement system for ESL learners. *Applied Measurement in Education*, 31(3):251–267. Publisher: Taylor & Francis.

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.

Wajdi Zaghouani. 2002. AUTO-ÉVAL : vers un modèle d'évaluation automatique des textes. In *Actes du colloque des étudiants en sciences du langage*, page 16, Montréal, Canada. Université du Québec à Montréal.

Haichao Zhu, Zekun Wang, Heng Zhang, Ming Liu, Sendong Zhao, and Bing Qin. 2021. Less is more: Domain adaptation with lottery ticket for reading comprehension. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1102–1113, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 19–27.

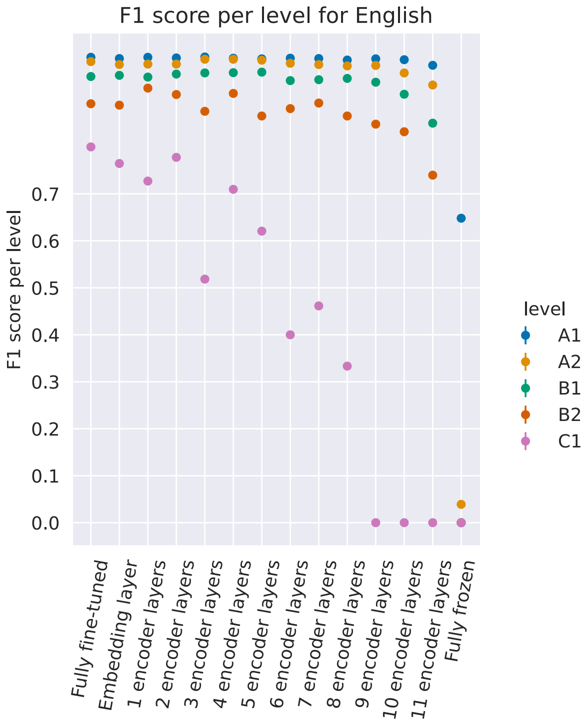

Figure 4: Performance per CEFR level when freezing different layers of the English model. Note that level A1 is the best performing one, while C1 is the worst.

A Performance Depending on the CEFR Level

In this appendix we present the figures for the F1 scores for the different languages. Figures 4, 5, and 6 show the effect of different degrees of fine-tuning of the BERT models across CEFR level in English, French, and Swedish, respectively.

B Detailed Results per Language

In this appendix we present tables with the usual metrics for each language. The ones based on adjacent accuracy are in Appendix C. Thus, Tables 5, 6, and 7 show the performance of different degrees of fine-tuning of the BERT models in English, French, and Swedish, respectively.

C Adjacent Metrics per Language

In this appendix we present tables with the metrics calculated using adjacent accuracy for each language. The ones based on standard accuracy are in Appendix B. Thus, Tables 8, 9, and 10 show the performance of different degrees of fine-tuning of the BERT models in English, French, and Swedish, respectively. Note that most of the experiments achieve very high results using these metrics.

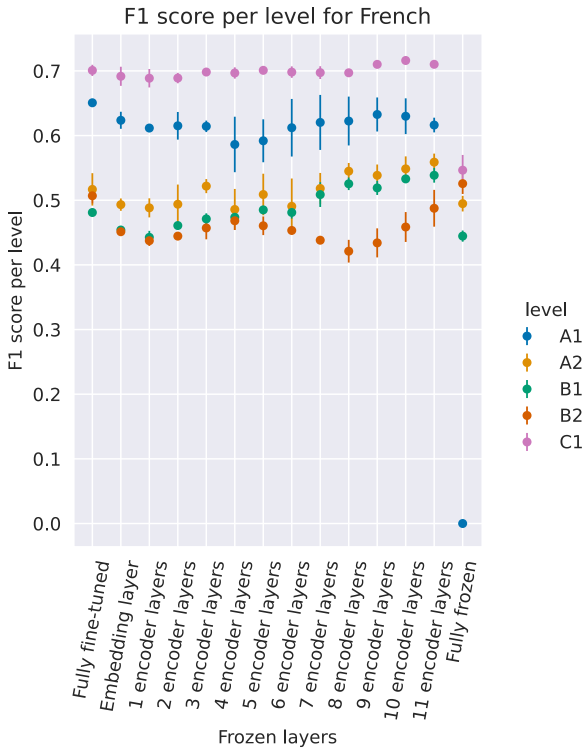

Figure 5: Performance per CEFR level when freezing different layers of the French model. Note that level C1 is the best performing one in general, followed by A1.

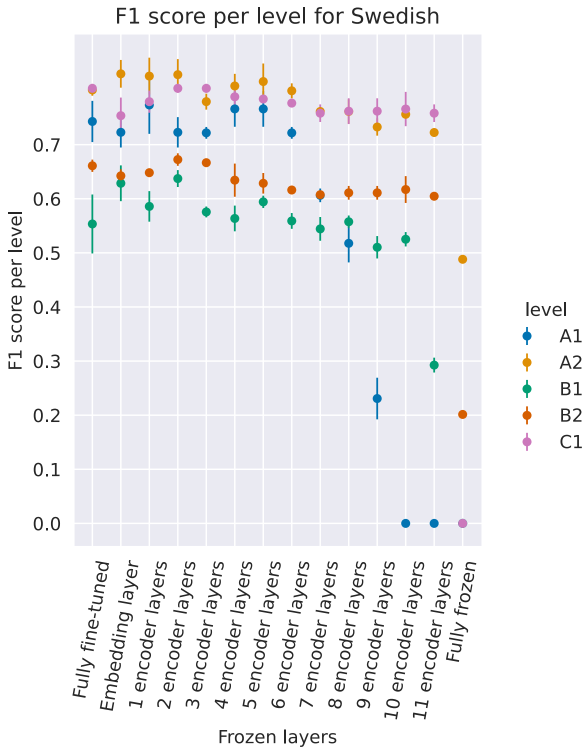

Figure 6: Performance per CEFR level when freezing different layers of the Swedish model. Note that levels A2 and C1 are the best performing ones in general, followed by A1.

Layers Frozen	**Accuracy**	**F1 macro**	**F1 weighted**
State-of-the-art (Schmalz and Brutti, 2021)	0.974	n/a	n/a
None	0.975 ± 0.000	0.923 ± 0.000	0.975 ± 0.000
All layers	0.475 ± 0.000	0.137 ± 0.000	0.319 ± 0.000
Embedding Layer	0.972 ± 0.000	0.914 ± 0.000	0.971 ± 0.000
1 Encoder Layer	0.974 ± 0.000	0.914 ± 0.000	0.974 ± 0.000
1 and 2	0.974 ± 0.000	0.922 ± 0.000	0.974 ± 0.000
1 to 3	0.975 ± 0.000	0.866 ± 0.000	0.974 ± 0.000
1 to 4	**0.977 ± 0.000**	**0.911 ± 0.000**	**0.977 ± 0.000**
1 to 5	0.973 ± 0.000	0.884 ± 0.000	0.972 ± 0.000
1 to 6	0.969 ± 0.000	0.838 ± 0.000	0.966 ± 0.000
1 to 7	0.969 ± 0.000	0.852 ± 0.000	0.967 ± 0.000
1 to 8	0.964 ± 0.000	0.820 ± 0.000	0.962 ± 0.000
1 to 9	0.962 ± 0.000	0.749 ± 0.000	0.957 ± 0.000
1 to 10	0.952 ± 0.000	0.737 ± 0.000	0.946 ± 0.000
1 to 11	0.924 ± 0.000	0.699 ± 0.000	0.919 ± 0.000

Table 5: Results of the various setups of English BERT model on the validation set using accuracy and macro and weighted F1. Note that the only result that outperforms a fully fine-tuned model was when freezing up to the fourth encoder layer. On top of that, the confidence interval was low enough for it to be considered practically zero.

Layers Frozen	**Accuracy**	**F1 macro**	**F1 weighted**
State-of-the-art (Wilkens et al., 2023)	0.57	n/a	0.56
None	0.560 ± 0.004	0.571 ± 0.003	0.555 ± 0.003
All layers	0.473 ± 0.005	0.402 ± 0.006	0.443 ± 0.005
Embedding Layer	0.533 ± 0.006	0.543 ± 0.004	0.526 ± 0.005
1 Encoder Layer	0.525 ± 0.011	0.534 ± 0.011	0.517 ± 0.011
1 and 2	0.533 ± 0.010	0.541 ± 0.012	0.524 ± 0.010
1 to 3	0.545 ± 0.003	0.553 ± 0.003	0.538 ± 0.002
1 to 4	0.538 ± 0.011	0.542 ± 0.014	0.529 ± 0.011
1 to 5	0.546 ± 0.008	0.549 ± 0.011	0.537 ± 0.008
1 to 6	0.542 ± 0.017	0.547 ± 0.020	0.532 ± 0.017
1 to 7	0.552 ± 0.018	0.557 ± 0.020	0.542 ± 0.018
1 to 8	0.559 ± 0.008	0.562 ± 0.010	0.548 ± 0.006
1 to 9	0.563 ± 0.006	0.567 ± 0.007	0.552 ± 0.004
1 to 10	0.573 ± 0.006	0.577 ± 0.007	0.564 ± 0.004
1 to 11	**0.578 ± 0.003**	**0.582 ± 0.002**	**0.572 ± 0.001**

Table 6: Results of the various setups of the French CamemBERT model on the validation set using accuracy and macro and weighted F1. Note that the best result on average is achieved when finetuning only the last encoder layer. More in general, finetuning the latter layers seems to lead to better results than also finetuning the earlier ones.

Layers Frozen	**Accuracy**	**F1 macro**	**F1 weighted**
State-of-the-art (Pilán et al., 2016)	0.18	0.16	0.23
None	0.727 ± 0.016	0.712 ± 0.021	0.722 ± 0.018
All layers	0.324 ± 0.004	0.138 ± 0.001	0.188 ± 0.001
Embedding Layer	0.731 ± 0.008	0.716 ± 0.008	0.727 ± 0.008
1 Encoder Layer	0.735 ± 0.020	0.723 ± 0.020	0.731 ± 0.019
1 and 2	**0.749** ± **0.012**	**0.733** ± **0.011**	**0.744** ± **0.011**
1 to 3	0.720 ± 0.008	0.710 ± 0.005	0.718 ± 0.006
1 to 4	0.724 ± 0.000	0.712 ± 0.003	0.720 ± 0.003
1 to 5	0.729 ± 0.012	0.718 ± 0.010	0.725 ± 0.010
1 to 6	0.710 ± 0.008	0.695 ± 0.005	0.705 ± 0.006
1 to 7	0.678 ± 0.008	0.656 ± 0.011	0.671 ± 0.009
1 to 8	0.673 ± 0.020	0.642 ± 0.021	0.664 ± 0.020
1 to 9	0.641 ± 0.016	0.569 ± 0.007	0.612 ± 0.011
1 to 10	0.649 ± 0.012	0.533 ± 0.014	0.596 ± 0.013
1 to 11	0.612 ± 0.000	0.476 ± 0.005	0.541 ± 0.004

Table 7: Results of the various setups of Swedish BERT model on the validation set using accuracy and macro and weighted F1. Note that the best result on average is achieved when finetuning the layers above the second encoder layer. Despite that, freezing some of the intermediate layers also leads to better results than those of the state-of-the-art.

Layers Frozen	**Adj. Accuracy**	**F1 macro**	**F1 weighted**
State-of-the-art (Schmalz and Brutti, 2021)	n/a	n/a	n/a
None	0.996 ± 0.000	0.997 ± 0.000	0.996 ± 0.000
All layers	0.799 ± 0.000	0.382 ± 0.000	0.721 ± 0.000
Embedding Layer	**0.998** ± **0.000**	**0.998** ± **0.000**	**0.998** ± **0.000**
1 Encoder Layer	0.996 ± 0.000	0.987 ± 0.000	0.996 ± 0.000
1 and 2	**0.998** ± **0.000**	**0.998** ± **0.000**	**0.998** ± **0.000**
1 to 3	0.996 ± 0.000	0.986 ± 0.000	0.996 ± 0.000
1 to 4	0.994 ± 0.000	0.984 ± 0.000	0.994 ± 0.000
1 to 5	0.994 ± 0.000	0.986 ± 0.000	0.994 ± 0.000
1 to 6	0.993 ± 0.000	0.964 ± 0.000	0.992 ± 0.000
1 to 7	0.993 ± 0.000	0.971 ± 0.000	0.993 ± 0.000
1 to 8	0.994 ± 0.000	0.988 ± 0.000	0.994 ± 0.000
1 to 9	0.995 ± 0.000	0.991 ± 0.000	0.995 ± 0.000
1 to 10	0.995 ± 0.000	0.990 ± 0.000	0.995 ± 0.000
1 to 11	0.996 ± 0.000	0.986 ± 0.000	0.996 ± 0.000

Table 8: Results of the various setups of English BERT model on the validation set using adjacent accuracy and the macro and weighted F1 scores that derive from it. Note that the best performance is achieved when freezing either just the embedding layer or by freezing up to the second encoder layer. This is the only model in which the best-performing does not match when using the usual accuracy and adjacent accuracy.

Layers Frozen	**Adj. Accuracy**	**F1 macro**	**F1 weighted**
State-of-the-art (Wilkens et al., 2023)	0.98	n/a	n/a
None	0.976 ± 0.002	0.976 ± 0.002	0.976 ± 0.002
All layers	0.952 ± 0.005	0.955 ± 0.005	0.952 ± 0.005
Embedding Layer	0.965 ± 0.001	0.966 ± 0.001	0.964 ± 0.001
1 Encoder Layer	0.958 ± 0.004	0.959 ± 0.003	0.958 ± 0.004
1 and 2	0.960 ± 0.002	0.961 ± 0.002	0.960 ± 0.002
1 to 3	0.965 ± 0.002	0.966 ± 0.002	0.965 ± 0.002
1 to 4	0.962 ± 0.001	0.963 ± 0.001	0.962 ± 0.001
1 to 5	0.960 ± 0.003	0.962 ± 0.002	0.960 ± 0.003
1 to 6	0.957 ± 0.004	0.958 ± 0.004	0.957 ± 0.004
1 to 7	0.960 ± 0.005	0.961 ± 0.005	0.960 ± 0.005
1 to 8	0.969 ± 0.002	0.970 ± 0.002	0.969 ± 0.002
1 to 9	0.972 ± 0.002	0.972 ± 0.002	0.972 ± 0.002
1 to 10	**0.976 ± 0.004**	**0.976 ± 0.003**	**0.976 ± 0.004**
1 to 11	**0.976 ± 0.002**	**0.976 ± 0.002**	**0.976 ± 0.002**

Table 9: Results of the various setups of French CamemBERT model on the validation set using adjacent accuracy and the macro and weighted F1 scores that derive from it. Note that the best result on average is achieved when finetuning either the final encoder layer or the final two.

Layers Frozen	**Adj. Accuracy**	**F1 macro**	**F1 weighted**
State-of-the-art (Pilán et al., 2016)	0.59	0.54	0.66
None	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
All layers	0.627 ± 0.012	0.585 ± 0.016	0.544 ± 0.015
Embedding Layer	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
1 - 10	1.000 ± 0.000	1.000 ± 0.000	1.000 ± 0.000
1 to 11	0.992 ± 0.004	0.993 ± 0.003	0.992 ± 0.004

Table 10: Results of the various setups of Swedish BERT model on the validation set using adjacent accuracy and the macro and weighted F1 scores that derive from it. Note that the best result on average is achieved when finetuning the layer above the third encoder one.