Comparing Traditional and LLM-based Approaches for Automated Scoring of Dutch Writing Products

Joni Kruijsbergen, Orphée De Clercq
LT^3, Ghent University
Groot-Brittanniëlaan 45, 9000 Ghent, Belgium
joni.kruijsbergen@ugent.be, orphee.declercq@ugent.be

Abstract

This research examines several traditional and recent approaches for automated grading of Dutch texts written by adolescent L1 speakers. We relied on a proprietary dataset comprising human-scored texts. Following recent paradigms in NLP research, we compared training a feature-based model to fine-tuning both mono- and multilingual BERT-based and generative large language models. The latter were also prompted directly in a zero-shot setting. The results reveal that the feature-based and BERT-based approaches are promising for the task at hand and even complementary, although there is still room for improvement. The error analysis demonstrates that the generative models do not only make more errors in classification, but that these error are also more problematic. We therefore conclude that especially generative LLMs are not directly employable in this educational context.

Keywords: Automated Essay Scoring, Writing Evaluation, Machine Learning, LLMs

1. Introduction

In today's knowledge-based society, the ability to write remains important in both professional and academic contexts (de Smet et al., 2012). However, writing entails highly demanding cognitive activity and requires the writer to be meticulous in their linguistic choices to ensure accurate delivery and reception of the intended message (Vasylets et al., 2017). Given these complexities, writing is often perceived as a challenging skill to master (Hayes and Flower, 1980; de Smet et al., 2012).

The media frequently report on declining proficiency skills, often fuelled by new results from large-scale studies such as OECD'S PIAAC (OECD, 2024). In Flanders, the region in Belgium in which the present study is set, writing skills have been shown to decrease among young and adolescent L1 writers (Snyder and a Dillow, 2014; De Smedt, 2019; Steunpunt Toetsontwikkeling, 2023). However, Belgium does not have a tradition of centralised testing which makes it difficult to assess the true depth of the problem. Additionally, chatbots driven by generative large language models (LLMs) are becoming an integral part of written language production, even for very young pupils (Picton and Clark, 2024). It is thus crucial to monitor writing skills to investigate the potential disruptive effect these technologies could have on writing in the future.

With the launch of centralised tests for reading and basic math skills among all pupils in the second year of secondary education in Flanders in 2024, a subgroup was also asked to write Dutch texts on a designated platform. This initiative aimed to collect written texts and explore the feasibility of assessing writing skills at a larger scale. For the present study, a subset of that data was scored by humans to investigate automated writing assessment.

Automated writing assessment has been thoroughly researched within NLP, mostly through the form of Automated Essay Scoring (AES) (Valenti et al., 2003; Ke and Ng, 2019; Li and Ng, 2024). Inspired by this prior work and following the current state of the art, we explore a data-driven approach. Despite the strong performance of pretrained LLMs on various NLP tasks (Min et al., 2023), recent work has shown that for AES, more traditional machine learning approaches are still on par (Benedetto et al., 2025). Furthermore, fine-tuning BERT-based LMs requires much less computational power than generative variants and performs comparable to feature-based (FB) methods (Mayfield and Black, 2020). Similarly to other NLP fields, the focus in AES research has mainly been on English, causing lower-resourced languages with less coverage in the pre-training data of LLMs to fall behind (Volodina et al., 2023). Specifically for LLMs it thus remains a challenge to improve their effectiveness on languages other than English.

This is exactly what will be researched in the present study. We compare a more traditional feature-based approach with fine-tuning both encoder- and decoder-based LLMs. To this end, all L1 Dutch written texts were reliably and holistically scored by humans using comparative judgment. Our research offers first insights into deploying LLMs for Dutch AES, and our results highlight the value of traditional approaches (FB and fine-tuning encoder-based LMs) for Dutch AES, especially when used in an educational setting.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 619–630
11–16 May 2026.

The remainder of this paper is structured as follows: Section 2 discusses related research in the field of AES. Section 3 provides information on the dataset and the proposed experimental setup, while Section 4 demonstrates and explains the overall evaluation strategy. In Section 5 the results are presented and discussed, accompanied by an in-depth error analysis. Section 6 concludes the paper while offering prospects for future research.

2. Related Work

Automated writing support systems have been extensively researched for decades, starting in the 1960s (Page, 1966). Today, such systems are prevalent as a result of advances in the fields of machine learning and NLP. Most well known are the proprietary systems for Automated Essay Scoring (AES), such as e-rater, which automatically grade written texts and are also the focus of the present study.

In their recent survey Li and Ng (2024) describe several categories of AES systems, among which machine learning (ML) and deep learning (DL) approaches. The more traditional ML approaches, such as a statistical classifier or regressor, use annotated corpora with extracted (linguistic) features to perform automated scoring. In these approaches, there is a strong focus on feature engineering and most systems are trained on a specific genre, referred to as within-domain scoring. DL approaches, on the other hand, are better at generalising and DL should thus work well for testing genres that are not part of the training set, i.e., cross-domain scoring (Li and Ng, 2024).

The state of the art in AES is often measured by using publicly available datasets such as ASAP — Automated Student Assessment Prize, which was introduced as part of a Kaggle competition. On that set, Mayfield and Black (2020) demonstrated that traditional FB models and fine-tuned BERT-based models perform similarly. Shin and Gierl (2021), on the other hand, showed that their approach using a convolutional neural network outperformed the more classical FB approach. Lagakis and Demetriadis (2021) concluded in their survey paper that the current state of the art is formed by combining fine-tuning BERT and curated features (2021). Regarding generative LLMs, Mizumoto and Eguchi (2023) researched GPT 3.5 for AES on the ETS Corpus of Non-Native Written English (Blanchard et al., 2014) and demonstrated weak agreement with the human reference scoring. More recently, Benedetto et al. (2025) showed that their feature-based baseline performed on par with all zero-shot generative models tested for AES on the KUPA-KEYS Corpus (Velentzas et al., 2024), and even outperformed some of them. They experimented with two proprietary and four open source models, one of which is the multilingual Llama 3 (Grattafiori et al., 2024). Based on those findings, we can reiterate Mayfield and Black's claims that using deep learning for AES is excessive, given that they provide only a slight improvement in performance and given the additional hardware and time requirements (2020). The latter two reasons are especially relevant considering the additional costs of current generative LLMs. Other issues are that these models occasionally fail to score all written texts and have a higher post-processing load (Benedetto et al., 2025).

The datasets used for benchmarking AES systems are often English learner essays, which results in a primary research focus on the English language and on the genre of argumentative essays. In order to gain more insight into the capabilities of the various AES approaches for other languages and genres, more research is crucial. The same applies to the broader field of automated writing support, where research on languages other than English is scarce, which is recognised as a problem (Benedetto et al., 2025). More recently, this has led to initiatives focusing specifically on multilingual approaches, such as the MultiGED (Volodina et al., 2023) and MultiGEC (Masciolini et al., 2025) shared tasks which dealt with the detection and correction of possible writing errors in 5 and 12 languages, respectively.

Not much work has been done on Dutch. Kruijsbergen et al. explored LLMs for grammatical error detection (2024) and in one prior study Dutch AES was explored by adapting ReaderBench to the Dutch language (Dascalu et al., 2017). In the latter study, a limited set of 173 technical reports were used, all written by Master Degree students who have relatively high writing skills. Although the research on Dutch is lacking, it can be considered a medium- to well-resourced language within NLP. For feature extraction, for example, one can rely on a linguistic processing tool such as T-Scan (Pander Maat et al., 2014) which is based on state-of-the-art NLP preprocessing software and was specifically designed for Dutch. At the same time, various Dutch (reference) corpora have been made available in the past through the STEVIN programme (Spyns and Odijk, 2013) and are incorporated as a curated part of the pre-training data for monolingual Dutch LLMs, such as RobBERT (Delobelle et al., 2020), GEITje Ultra (Vanroy, 2024) and ChocoLlama (Meeus et al., 2024). However, when it comes to automated writing support the bottleneck is often consistently labelled authentic and representative data (Kruijsbergen et al., 2024).

With this research, we want to demonstrate the possibilities of AES for Dutch, relying on traditional FB methods and newer approaches using LLMs.

As this project is set in the framework of centralised tests, we have to take into account the educational environment, which requires both high model interpretability and practical feasibility. By doing this research and demonstrating the results for Dutch, we hope to help open up the field to more research on automated writing support in general, and especially for languages other than English.

3. Methodology

3.1. Data

The dataset used for this study originates from a pilot project that took place within the first iteration of centralised tests that were taken by all Flemish pupils in the second year of secondary education in 2024. For this pilot, 1,426 written texts were collected from a random sample of pupils. The writing prompt instructed the pupils to write two articles for their school newspaper, aimed at their fellow students: (1) an informative text in which they had to write about their dream job, and (2) a persuasive text in which they had to convince fellow pupils to try out their hobby. For both articles the pupils were prompted to write a text of around 200 words. We found that the average word count in our dataset amounted to 195 words, with a standard deviation of 57 words.

For the evaluation of the written texts we opted for holistic scoring through comparative judgment. Advantages of comparative judgement are that the process is much faster than scoring using an analytic rubric and that experts are not needed to grade the written texts, in contrast to analytic scoring (Bouwer et al., 2023). Furthermore, comparative judgement has proven to be on par with analytic scoring outcomes, provided that there are enough assessors and that a reliability level of minimum 0.70 is reached (van Daal et al., 2022). To operationalise this, we relied on the Comproved tool. Interestingly, this tool allows for setting a reliability threshold beforehand, which helps to delineate how many human assessors are needed, by providing an estimate beforehand and showing the reliability score during scoring. In line with the work of Bouwer et al. (2023) – which considers 0.70 *acceptable*, but 0.80 *good* rater agreement – we set this threshold at 0.80 and recruited the required number of assessors, namely 11.

An example of the Comproved assessment interface is presented in Figure 1. The informative and persuasive texts were assessed in two separate runs, after which a separate ranking for each genre was obtained (see Figure 2 for the ranking of the persuasive texts). Since the aim is eventually to grade whether or not a text passes the final objectives of Dutch writing comprehension, we included

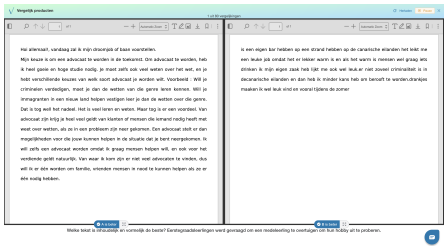

Figure 1: Example of two informative texts in the Comproved web application. The texts describe pupils' dream jobs, and assessors are asked to choose the best text in terms of content and form.

in both sets a benchmark text selected by experts which can be considered as *just sufficient*. This allowed us to appoint binary *pass* and *fail* labels to all evaluated written texts in the ranking, i.e. all texts that are lower in the ranking than the benchmark text receive *fail*, and all other texts *pass*.

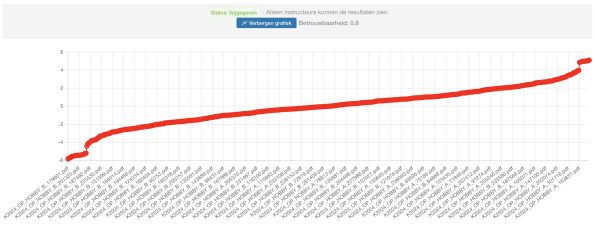

Figure 2: Ranking of the persuasive texts through comparative judgement in the Comproved web application.

3.2. Experimental setup

As explained in Section 2, our objective is to investigate the feasibility of creating a data-driven grading system to assign a pass or fail to Dutch writing products, following three machine learning paradigms: traditional feature-based machine learning, fine-tuning BERT-based LMs and prompting present-day generative LLMs. Regarding the LLMs, we opted to work with both monolingual and multilingual variants.

For the FB experiments, linguistic characteristics were automatically derived from the written texts in the form of summative features. These features most notably capture aspects of vocabulary, syntax and cohesion and have proven good predictors of writing quality (Crossley, 2020). To this end, we relied on T-Scan (Pander Maat et al., 2014), which has been specifically designed and optimised to process Dutch texts. T-Scan extracts more than 400 features in its current state, with features concerning lexical complexity, sentence complexity, referential cohesion and lexical diversity, lexical semantics and personal style (Pander Maat et al., 2014).

For the **feature-based classifications**, we decided on a random forest classifier. This model came out on top in pilot experiments that were conducted on a similar dataset and in which random forests were compared to plain decision trees, gradient boosting and support vector machines. For the implementation of the random forest classifier, we made use of the open-source library scikit-learn. The features were standardised before training. Because of the high number of T-Scan features, we also performed Lasso regressions to select only the features with non-zero coefficients after which the classifier was retrained. Reducing the number of features can provide more insight into which features are impactful, therefore improving model interpretability. To this end, scikit-learn was consulted again. The optimal regularisation strength for Lasso was calculated on the same in-house pilot set and set to 0.005. For all experiments, the random state was set to 42 for reproducibility.

Regarding the fine-tuning of **BERT-based LMs**, we opted for the monolingual Dutch RobBERT (Delobelle et al., 2020) and multilingual XLM-RoBERTa (Conneau et al., 2020) models. This choice was made as similar work on Dutch discussed the relevance of comparing monolingual to multilingual models and demonstrated better results for RobBERT and XLM-RoBERTa than for other BERT-based models for low-proficient writing (Kruijsbergen et al., 2024). For fine-tuning, we used HuggingFace's transformers library. The training and validation batch sizes were reduced to 4 due to memory constraints, but all other parameters were left as default.

For the **generative LLMs**, fine-tuning and zero-shot approaches were used on open source models. Similarly to the BERT-based models, we compared monolingual Dutch LLMs with a multilingual counterpart. Moreover, we chose to prompt the multilingual model in both Dutch and English as previous research demonstrated mixed results as to which prompting language resulted in better output for Dutch (Kruijsbergen et al., 2024). Regarding the monolingual model, we relied on the two generative models available for Dutch, being GEITje 7B Ultra (Vanroy, 2024) and ChocoLlama (Meeus et al., 2024). For the multilingual model, Llama 3 8B instruct (Grattafiori et al., 2024) was chosen.

For zero-shotting and fine-tuning the generative LLMs, we accessed the models via the AutoModelForCausalLM class of the HuggingFace transformers library. We used parameter-efficient fine-tuning with QLoRa to avoid adjusting all the model weights (Dettmers et al., 2023). The training and prediction parameters can be found as Appendix A. For the zero-shot experiments, we directly prompted the LLM, and for fine-tuning, we combined prompting with the gold standard examples from the training set. The complete prompts are available in Appendix B.

4. Evaluation

With these experiments, we investigated the automatic scoring of texts written by young native Dutch speakers and graded by human assessors. The task at hand was binary classification, where each text is automatically assigned a label (*pass* or *fail*).

	Informative	**Persuasive**	**Combined**
pass	453	429	882
fail	258	286	544
total	711	715	1426

Table 1: Number of passed and failed texts for the informative and the persuasive texts, as well as for the combined texts.

Although the two text types were evaluated separately, we can merge both sets into one combined dataset after assigning each text a label. Table 1 presents the informative, persuasive and combined sets. As can be observed, there is slight imbalance, with more passes than fails.

To evaluate, we report precision, recall and F-score. We opt for the balanced F_1 score, which is a harmonic mean between precision and recall. In high-stakes environments it has been demonstrated to be more detrimental that a text receives the label *fail* when in fact it is good enough to pass (Machin et al., 2020), rendering recall more important than precision. However, since our case is low-stakes, we choose the harmonic mean, but will particularly focus on those misclassifications in the error analysis (Section 5.3). The reported F_1 scores are macro-averaged, precision and recall are reported per class, and a majority baseline each time predicting the dominant (*pass*) class is added.

In order to ensure robustness, a cross-validation setup is often implemented in AES research. However, running generative models is computationally expensive, especially in terms of sustainability (Berthelot et al., 2025). Furthermore, the high variability in their output often requires additional post-processing (Kruijsbergen et al., 2024; Benedetto et al., 2025). Therefore, we decided to perform the experiments in two settings. Setting 1 compares feature-based and BERT-based models in a ten-fold cross-validation setup (Section 4.1), and Setting 2 compares the best versions of those models with the generative AI models using a fixed training, development and test split (Section 4.2).

4.1. Cross-validation setting

For the experiments in Setting 1, all data are first used in a ten-fold cross-validation setup, for the *informative* and *persuasive* sets separately. As previously shown (Table 1), there is a slight imbalance in our dataset towards the positives. For the informative, persuasive and combined sets, approximately 60% have a positive label and nearly 40% a negative label.

Over the ten folds, data is split automatically into an 80/10/10 training, development and test split, using scikit-learn's train_test_split with random state 42. To calculate the scores, we average over all ten folds computing average precision, recall and F_1 score.

This setup allows us to check the robustness of FB and DL approaches. For the best models of both approaches, we also investigate their effectiveness at cross-domain scoring by training on the full informative set and testing on folds of the persuasive set in a pseudo-cross-validation setup, and vice versa. Testing this will verify whether Li and Ng's claim (2024) that deep learning is better at cross-domain evaluation also holds for Dutch.

4.2. Fixed Split + LLM Evaluation

In Setting 2, generative models are also compared to the best performing FB and DL. Here only the combined set is used, to ensure enough data for fine-tuning the generative models, and split into an approximate 70/20/10 training, development and test split (see Table 2 for the label distribution in the three splits). Training and validation splits were used for the development of the feature-based and fine-tuned models, which were then evaluated on the held-out test set. Since zero-shot experiments do not require any training data, the held-out test split was directly used to evaluate the zero-shot prompting approach. As Table 2 shows, the distribution in the test split is slightly less balanced than in the training and development splits.

train		**dev**		**test**	
pass	fail	pass	fail	pass	fail
636	394	152	103	94	47

Table 2: Number of passed and failed texts in the training, development and test splits for the combined texts.

5. Results & Discussion

This section presents and discusses the results of the various models for the automated classification of written texts written by pupils in the second year of secondary education. A distinction is made between Section 5.1, which tests the robustness of the BERT-based and feature-based models by performing cross-validation and cross-domain scoring, and Section 5.2, which compares the BERT-based LMs and more traditional feature-based classifiers to generative LLMs through one run with fixed data splits. The section ends with a thorough discussion and error analysis that compares incorrect predictions from the best models in the fixed setup (Section 5.3).

5.1. Results Cross-validation

The results in Table 3 show consistent scores for both feature-based (FB) approaches and when fine-tuning RobBERT, with F_1 scores around 80%. If we look at the FB results we notice that feature selection through Lasso yields the best results. More pronounced is the drop in performance for XLM-RoBERTa from the informative to the persuasive genre (F_1 of 80.44% vs. 65.06%). The fine-tuning of the monolingual RobBERT LLM, on the other hand, leads to the overall highest F_1 score for both genres.

Table 3 also shows the results for cross-domain scoring on the two best models: FB with Lasso and DL with RobBERT. For the round where we trained on the persuasive texts and tested on the informative texts, we see a rise for precision of the texts with label *pass*, and a drop regarding recall, as well as a drop regarding precision of the texts with label *fail* and a rise in recall score for that label. F_1 score is slightly lower than for within-domain testing. For the round where we tested on the persuasive texts, the opposite holds true: a small decrease in precision scores for *pass* and recall scores for *fail*, and a higher score for recall *pass* and precision *fail*. Overall, the F_1 results on the persuasive test are very similar for FB and DL, and on the informative set, RobBERT performs slightly better. This indicates that both FB and DL approaches work for cross-domain evaluation with the presented genres. Their similar scores also refute the claim of (Li and Ng, 2024) that DL is inherently better at cross-domain evaluation, at least for Dutch and with these genres and writing tasks.

Regarding within-domain evaluation, we also examined the combined set (informative + persuasive texts). Those scores are in line with the scores for the informative and persuasive sets, with the best models scoring 78.41% and 82.17% F_1-scores for the FB model with Lasso and the RobBERT model, respectively. The fact that most models score similarly on the informative and combined sets is in line with the expectations. On the one hand, doubling the training data ensures more examples for the model, which should render classification easier.

		Informative					Persuasive				
		P(p)	R(p)	P(f)	R(f)	F1	P(p)	R(p)	P(f)	R(f)	F1
Within-domain	**FB**	0.822	0.893	0.776	0.664	0.783	0.808	0.872	0.782	0.690	0.784
	FB Lasso	**0.835**	0.878	0.767	**0.704**	0.794	**0.810**	0.865	0.776	**0.700**	0.784
	RobBERT	0.834	**0.945**	**0.879**	0.665	**0.818**	0.794	0.938	0.876	0.637	**0.797**
	XLM	0.830	0.926	0.837	0.667	0.804	0.727	0.939	0.582	0.432	0.651
Cross-domain	**FB Lasso**	0.875	0.774	0.667	0.805	0.774	0.749	0.921	0.821	0.538	0.737
	RobBERT	0.911	0.795	0.695	0.852	0.806	0.732	**0.968**	**0.898**	0.465	0.721

Table 3: Results for the cross-validation experiments (within-domain) on the informative and persuasive sets with the feature-based (FB) base and Lasso models and the fine-tuned deep learning (DL) models RobBERT and XLM-RoBERTa, supplemented with the scores on the cross-domain experiments for the best feature-based (with Lasso) and deep learning (RobBERT) model. Results are provided regarding precision P and recall R for the classes pass (p) and fail (f), and macro-averaged F_1 score.

On the other hand, in terms of content and writing objective, writing an informative text is less complex than writing a persuasive text, which renders informative texts easier to grade based on more superficial features. This is also a possible explanation for the higher scores on the informative test sets in the cross-domain evaluation setup.

5.2. Results Fixed Split

Table 4 shows the results for the best FB and BERT-based models and the best versions of the three generative LLMs: the fine-tuned setting produced the best results for GEITje and ChocoLlama, and the zero-shot setting prompted in Dutch for Llama. For a full overview of the results, see Appendix C.

The F_1 scores for the FB and BERT-based models are considerably high, with an F_1 of 81.75% for FB and 87.09% for RobBERT.

In terms of the generative LLMs, there is a significant difference between the models. Table 4 shows that the best version of ChocoLlama, for example, barely outperforms the baseline with an F_1 score of 47.30%. Fine-tuned GEITje scores better (61.87%) but even the best Llama version scores much lower than the feature-based (difference of more than 5 percentage points (pp)) and BERT-based models (difference of more than 10pp).

Interestingly, for GEITje and ChocoLlama, zero-shot results are worse than their fine-tuning scores (F_1 of 44.79% and 44.29%, respectively) and barely better than the baseline. The increases in fine-tuning scores demonstrate that fine-tuning a generative model can have a positive effect on AES results. However, for Llama, the fine-tuning results show that the pre-training data can outweigh the effect of fine-tuning, given that the zero-shot models actually perform better for both the Dutch and English prompt. Regarding the English prompt, the scores remain low with 58.76% (ZS) and 49.93% (FT), but when prompted in Dutch the Llama model scores considerably better, especially in the zero-shot setting (76.74% ZS, 70.71% F_1).

5.3. Error analysis

To provide more insight into the errors made by the models, we took a closer look at the incorrect predictions made on the test split from setting 2, as we can easily compare those for all model types. We did this for the models with the best F_1 score per type, namely the feature-based model with Lasso (81.75%), the RobBERT model for the BERT-based LM (87.09%), and the Dutch-prompted zero-shot Llama for the generative LLMs (76.74%).

Firstly, we wanted to find out whether the model predictions were complementary. To this purpose, we created an *aggregated model* by using the majority vote between the three models (FB with Lasso, RobBERT and Llama): if at least two models predicted a certain label, the aggregated model also predicts that label (See row *Aggregated (1)* in Table 5). The results show a slight increase in terms of precision, but mostly for recall and F_1 score, compared to the best scoring model (RobBERT). This confirms that the models are at least partially complementary in their predictions.

Since the scores of the generative models were overall much worse than the feature- and BERT-based models, we also wanted to investigate how complementary these latter two models are and whether the generative model is actually necessary. To do so, we checked the cases in which the FB and BERT-based model disagreed on the prediction (13 cases, or 9.22%). If the models agreed, we chose that given label and if they disagreed, we chose the majority label in the whole dataset, namely *pass*. The results (See row *FB + RobBERT (2)* in Table 5) indicate a slightly lower precision score than for the RobBERT model, but an even higher recall score than the aggregated model, as well as the highest F_1 score - admittedly with a small margin compared to the aggregated model. However, this indicates that similar high results can be obtained without the need for a generative LLM.

As is the case for human grading, some texts are more difficult to grade than others, which is particularly the case with texts that are neither very

	P(p)	R(p)	P(f)	R(f)	F1
FB Lasso	0.882	0.872	0.75	0.766	0.818
RobBERT	**0.906**	**0.926**	**0.844**	**0.809**	**0.871**
GEITje	0.753	0.713	0.481	0.532	0.619
ChocoLlama	0.763	0.309	0.369	**0.809**	0.473
Llama	0.842	0.851	0.696	0.681	0.767
Baseline	0.667	1	0	0	0.4

Table 4: Scores on the fixed test split for the best feature-based models and the best BERT-based and generative LLMs, as well as the baseline. For GEITje and ChocoLlama, this is the Dutch prompted fine-tuned setting, for Llama it is the Dutch prompted zero-shot setting. Results are provided regarding precision P and recall R, for classes pass (p) and fail (f), and macro-averaged F_1 score.

	P	R	F1
FB Lasso	0.816	0.819	0.818
RobBERT	0.875	0.867	0.871
Llama	0.769	0.766	0.767
Aggregated (1)	**0.887**	0.915	0.901
FB + RobBERT (2)	0.871	**0.936**	**0.903**

Table 5: Scores on the fixed test split for the best models, supplemented with the scores of (1) an aggregated model using majority vote between the three best models for feature-based, BERT-based and generative LLM-based experiments, and (2) a model outputting the predicted label if the best feature-based and BERT-based model agree, and otherwise the majority label.

good nor bad. In those cases, it can also be difficult for an expert to assess text quality (Wolfe et al., 2016). Therefore, we examined the incorrectly labelled texts for all three models, i.e., texts that had a *pass* as gold label but received a *fail* by the machine, or vice versa. We manually checked whether those incorrectly labelled texts could be considered good, bad or mediocre as part of the error analysis. This was possible using the original scores from the comparative judgement output. Those scores range from 0 to 1, with a score closer to 1 representing a better text.

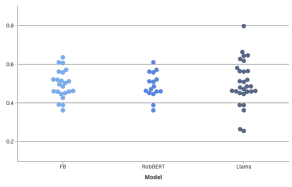

Figure 3: Distribution of the incorrect predictions for the best Feature-based (FB), RobBERT and Llama model.

The visualisation in Figure 3 clearly indicates that the incorrectly labelled texts from the RobBERT- and feature-based model are situated around the centre, with the distribution of the RobBERT model being slightly more narrow. Interestingly, upon closer inspection, the incorrect labels from RobBERT that reach the furthest from the middle, correspond to texts which also received an incorrect label from at least one of the other models. Overall, a clear case can be made for the *grey area* or middle ground where texts are known to be more difficult to grade. This means that really high- and low-quality texts are at least scored correctly by both the feature-based and RobBERT models. However, that is not the case for the Llama model, which has more problematic incorrect predictions: two bad texts that received a *pass* and one high scoring text that received a *fail*. This underlines that generative LLMs are not ready for direct implementation in educational contexts.

Lastly, we specifically considered texts that were misclassified by all three models: one text that was labelled *fail* by all models, but was actually a *pass* and seven texts that were labelled *pass*, when they actually failed for the human assessors. With regard to the latter, most texts can be considered adequate in terms of spelling and grammar, but lacking in terms of content and writing objective. That means that the aspects of content and writing objective seem more important to human assessors than to the models. For the former, we also regarded the three texts mentioned above for which the RobBERT and FB model erroneously predicted the label *fail* and Llama predicted the correct label *pass*, since failing someone who deserves to pass can be detrimental, especially in high-stakes environments (Machin et al., 2020). Although less pronounced, those texts showed a reverse trend: they were mostly good in terms of content and writing objective, but looked worse in terms of spelling, grammar and punctuation.

6. Conclusion

With this paper, we aimed to compare traditional feature-based models to current state-of-the-art large language models on the task of automatically assessing Dutch written texts, targeting pupils in

the second year of secondary education in Flanders.

More specifically, we investigated more traditional machine learning approaches, such as using a feature-based random forest model, as well as more recent deep learning approaches, such as fine-tuning BERT-based and generative large language models. For both deep learning approaches, we examined the possibilities of monolingual Dutch (RobBERT, GEITje & ChocoLlama) and multilingual (XLM-RoBERTa & Llama) open-source LLMs, since these have demonstrated state-of-the-art performance in similar automated writing support tasks.

To this purpose written texts of two genres were scored through comparative judgement and used to train and compare the three approaches for within-domain evaluation. The results indicate that the feature-based and fine-tuned RobBERT models are most consistent, with averages of around 80% F_1 score. Furthermore, the error analysis revealed that incorrect predictions for these models were generally texts in the *grey area*, meaning that they are more difficult to predict correctly than really good or bad texts. Moreover, we tested the possibilities of employing the FB-based and RobBERT models for cross-domain evaluation, which resulted in comparable scores for both approaches. Overall, these models produced good results, although there is room for improvement, since nearly 20% of the predictions are still consistently wrong.

The generative models were overall much less consistent, barely able to outperform the baseline with GEITje, ChocoLlama and Llama when relying on Dutch prompts. Even in the best approach, namely English-prompted Llama in a zero-shot setting, results are still far from the ones from the feature-based and BERT-based models. Moreover, the error analysis indicates that 1) the generative model is not necessary to obtain the best results, which can be achieved by aggregating the best feature-based model with RobBERT, and that 2) the generative model makes more problematic incorrect predictions. Some other remaining issues with generative LLMs are that they have a much higher computational cost than the other approaches and that the post-processing load is much higher.

This highlights that generative models are not suitable for Dutch AES for young learners in their current state. Furthermore, we conclude that although the other approaches yielded promising results, these serve as a first proof-of-concept but are not ready for direct implementation in the educational field without further research and development. A first future research direction will be to train and test the systems on larger and more diverse datasets and genres and to explore other forms of manual scoring, such as analytical scoring by means of a rubric.

Limitations

Although the dataset used is representative for the target audience, the test split of 141 texts is not very large. This makes it difficult to draw grounded conclusions, yet it still provides first insight into the models' capabilities for Dutch AES for these young L1 learners. In future work, we would like to expand annotations of the rest of the original set, so that we have more texts to work with.

An additional limitation of the dataset is that it is currently only for in-house use and can therefore not be shared. We are aware that this reduces the reproducibility of the results.

Next, although we were able to perform an in-depth error analysis, we believe it would be easier to find systematic and objective grounds for a similar analysis if the texts are also enriched with analytic scoring through a rubric. That can help to provide more insight into the specific criteria that the models are able to grasp and therefore help interpretability, which is crucial when models are intended for an educational context. Additionally, this would enable an evaluation of the models' ability to provide more fine-grained assessment beyond a binary pass-fail label. This is an avenue of future work.

Another issue with the present setup is that it is difficult to compare our scores directly to other available scores for this task for several reasons. Firstly, those other experiments are often performed on texts written in English, and the background of the pupils can vary. Next, the text types differ over datasets, and the context can be really different (e.g., high-stakes vs. low-stakes environment).

Acknowledgements

We would like to thank the anonymous reviewers for their valuable insights. Additionally, our thank goes out to Steunpunt Centrale Toetsen for providing us with the in-house dataset and for supporting this work. This work was also supported by Ghent University under grant BOF.STG.2022.0012.01 and by the Research Foundation–Flanders under grant number FWO.SPB.2025.0019.01.

Bibliographical References

Luca Benedetto, Gabrielle Gaudeau, Andrew Caines, and Paula Buttery. 2025. Assessing how accurately large language models encode and

apply the common European framework of reference for languages. *Computers and Education: Artificial Intelligence*, 8:100353.

Adrien Berthelot, Eddy Caron, Mathilde Jay, and Laurent Lefèvre. 2025. Understanding the environmental impact of generative ai services. *Commun. ACM*, 68(7):46–53.

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2014. ETS Corpus of Non-Native Written English.

Renske Bouwer, Monica Koster, and Huub van den Bergh. 2023. Benchmark rating procedure, best of both worlds? Comparing procedures to rate text quality in a reliable and valid manner. *Assessment in Education: Principles, Policy & Practice*, 30(3-4):302–319.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Scott A. Crossley. 2020. Linguistic features in writing quality and development: An overview. *Journal of Writing Research*, 11(3):415–443.

Mihai Dascalu, Wim Westera, Stefan Ruseti, Stefan Trausan-Matu, and Hub Kurvers. 2017. ReaderBench Learns Dutch: Building a Comprehensive Automated Essay Scoring System for Dutch Language. In *Artificial Intelligence in Education*, pages 52–63, Cham. Springer International Publishing.

Fien De Smedt. 2019. *Cognitive and motivational challenges in writing : the impact of explicit instruction and peer-assisted writing in upper-elementary grades*. Ph.D. thesis, Ghent University.

Milou J.R. de Smet, Saskia Brand-Gruwel, Hein Broekkamp, and Paul A. Kirschner. 2012. Write between the lines: Electronic outlining and the organization of text ideas. *Computers in Human Behavior*, 28(6):2107–2116.

Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. RobBERT: a Dutch RoBERTa-based Language Model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3255–3265, Online. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLORA: efficient finetuning of quantized LLMs. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, ..., and Zhiyu Ma. 2024. The Llama 3 Herd of Models.

John R. Hayes and Linda Flower. 1980. Identifying the Organization of Writing Processes. In Lee W. Gregg and Erwin R. Steinberg, editors, *Cognitive Processes in Writing: An Interdisciplinary Approach*, pages 3–30. Lawrence Erlbaum, Hillsdale, NJ.

Zixuan Ke and Vincent Ng. 2019. Automated Essay Scoring: A Survey of the State of the Art. In *Proceedings of the Twenty-Eight International Joint Conference on Artificial Intelligence*, pages 6300–6308. International Joint Conferences on Artificial Intelligence Organization.

Joni Kruijsbergen, Serafina Van Geertruyen, Véronique Hoste, and Orphée De Clercq. 2024. Exploring LLMs' Capabilities for Error Detection in Dutch L1 and L2 Writing Products. *Computational Linguistics in the Netherlands Journal*, 13:173–191.

Paraskevas Lagakis and Stavros Demetriadis. 2021. Automated essay scoring: A review of the field. In *2021 International Conference on Computer, Information and Telecommunication Systems (CITS)*, pages 1–6.

Shengjie Li and Vincent Ng. 2024. Automated Essay Scoring: A Reflection on the State of the Art. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17876–17888, Miami, Florida, USA. Association for Computational Linguistics.

Stephen Machin, Sandra McNally, and Jenifer Ruiz-Valenzuela. 2020. Entry through the narrow door: The costs of just failing high stakes exams. *Journal of Public Economics*, 190:104224.

Arianna Masciolini, Andrew Caines, Orphée De Clercq, Joni Kruijsbergen, Murathan Kurfalı, Ricardo Muñoz Sánchez, Elena Volodina, and Robert Östling. 2025. The MultiGEC-2025 shared task on multilingual grammatical error correction at NLP4CALL. In *Proceedings of the*

14th Workshop on Natural Language Processing for Computer Assisted Language Learning, pages 1–33, Tallinn, Estonia. University of Tartu Library.

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.

Matthieu Meeus, Anthony Rathé, François Remy, Pieter Delobelle, Jens-Joris Decorte, and Thomas Demeester. 2024. ChocoLlama: Lessons Learned From Teaching Llamas Dutch.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2023. Recent Advances in Natural Language Processing via Large Pre-Trained Language Models: A Survey. *Association for Computing Machinery*, 56(2).

Atsushi Mizumoto and Masaki Eguchi. 2023. Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, 2(2):100050.

OECD. 2024. Do Adults Have the Skills They Need to Thrive in a Changing World?: Survey of Adult Skills 2023.

Ellis B. Page. 1966. The Imminence of... Grading Essays by Computer. *The Phi Delta Kappan*, 47(5):238–243.

Henk Pander Maat, Rogier Kraf, Antal van den Bosch, Nick Dekker, Maarten van Gompel, Suzanne Kleijn, Ted Sanders, and Ko van der Sloot. 2014. T-Scan: a new tool for analyzing Dutch text. *Computational Linguistics in The Netherlands Journal*, 4:53–74.

Irene Picton and Christina Clark. 2024. Children and Young People's Use of Generative AI to Support Literacy in 2024. Accessed: 2025-04-17.

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.

Thomas D Snyder and Sally a Dillow. 2014. *Digest of education statistics 2012*. Government Printing Office.

Peter Spyns and Jan Odijk. 2013. *Essential Speech and Language Technology for Dutch: Results by the STEVIN-programme*. Springer Berlin, Heidelberg.

Steunpunt Toetsontwikkeling. 2023. Peiling Nederlands in de eerste graad secundair onderwijs 2022.

Salvatore Valenti, Francesca Neri, and Alessandro Cucchiarelli. 2003. An Overview of Current Research on Automated Essay Grading. *Journal of Information Technology Education: Research*, 2(1):319–330.

Tine van Daal, Marije Lesterhuis, Sven De Maeyer, and Renske Bouwer. 2022. Editorial: Validity, reliability and efficiency of comparative judgement to assess student work. *Frontiers in Education*, 7:1100095.

Bram Vanroy. 2024. GEITje 7B Ultra: A Conversational Model for Dutch.

Olena Vasylets, Roger Gilabert, and Rosa M. Manchón. 2017. The Effects of Mode and Task Complexity on Second Language Production. *Language Learning*, 67(2):394–430.

Georgios Velentzas, Andrew Caines, Rita Borgo, Erin Pacquetet, Clive Hamilton, Taylor Arnold, Diane Nicholls, Paula Buttery, Thomas Gaillat, Nicolas Ballier, and Helen Yannakoudakis. 2024. Logging Keystrokes in Writing by English Learners. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10725–10746, Torino, Italia. ELRA and ICCL.

Elena Volodina, Christopher Bryant, Andrew Caines, Orphée De Clercq, Jennifer-Carmen Frey, Elizaveta Ershova, Alexandr Rosen, and Olga Vinogradova. 2023. MultiGED-2023 shared task at NLP4CALL: Multilingual grammatical error detection. In *Proceedings of the 12th Workshop on NLP for Computer Assisted Language Learning*, pages 1–16, Tórshavn, Faroe Islands. LiU Electronic Press.

Edward W. Wolfe, Tian Song, and Hong Jiao. 2016. Features of difficult-to-score essays. *Assessing Writing*, 27:1–10.

Appendices

A. Model parameters

Fine-tuning	
Max Sequence Length	1024
Lora Alpha	16
Ranks	128
Learning Rate	5e-5
Batch Size	4
Training Epochs	1
Prediction	
Max Sequence Length	150
Top K	10
Returned Sequence	1

Table 6: Model parameters for the fine-tuning and prediction phases of the generative LLMs.

B. Prompts

B.1. English

System prompt:

```
You are a Dutch teacher in the
second year of secondary education
in Flanders.  That makes you an
expert in grading writing products.
The exercise was to inform the
other students of their dream job
or to convince them to try out their
hobby, both of which in Dutch.  The
possible labels are:  'The quality
of this text is insufficient and can
therefore not get a passing grade.'
and 'This text is good enough to
pass for this exercise.'
```

Prompt template:

```
{user} Is this text which is written
by a 13 year old good enough to get
a passing grade for the writing
exercise of informing the other
students of their dream job or to
convincing them to try out their
hobby?  Do not give any explanations,
limit yourself to the label.
### Text:  {placeholder for input
text}
{assistant} ### Label:  {placeholder
for label}
```

B.2. Dutch

System prompt:

```
Je bent een leerkracht Nederlands in
het tweede middelbaar in Vlaanderen.
Dat maakt jou een expert in het
beoordelen van schrijfproducten.
De schrijfopdracht was om hun
medeleerlingen te informeren over
hun droomjob of te overtuigen om
hun hobby uit te proberen.  De
mogelijke labels zijn:  'Deze tekst
is ondermaats en kan niet slagen
op de schrijfopdracht.'  en 'Deze
tekst is op het juiste niveau dus
goed genoeg om te slagen op de
schrijfopdracht.'
```

Prompt template:

```
{user} Is deze tekst van een
leerling van 13 jaar goed genoeg
om te slagen op de schrijfopdracht
om medeleerlingen te informeren over
hun droomjob of te overtuigen om hun
hobby uit te proberen?  Geef geen
uitleg waarom, maar beperk je tot
alleen het label.
### Text:  {placeholder for input
text}
{assistant} ### Label:  {placeholder
for label}
```

C. Full results fixed data split

	FB		DL		GenAI							
	Base	Lasso	RobBERT	XLM	GEITje		ChocoLlama		Llama			
					ZS	FT	ZS	FT	ZS NL	FT NL	ZS EN	FT EN
P(p)	0.872	0.882	**0.906**	0.906	0.625	0.753	0.623	0.763	0.842	0.875	0.715	0.853
R(p)	0.872	0.872	0.926	0.819	0.538	0.712	0.511	0.309	0.851	0.670	**0.989**	0.309
P(f)	0.745	0.75	0.844	0.696	0.283	0.481	0.281	0.369	0.696	0.551	**0.909**	0.393
R(f)	0.745	0.766	0.809	0.83	0.362	0.532	0.383	0.809	0.681	0.809	0.213	**0.894**
F1	0.809	0.818	**0.871**	0.809	0.448	0.619	0.443	0.473	0.767	0.707	0.588	0.499

Table 7: Results for the experiments with the fixed data splits, using the feature-based (FB) basic and Lasso models, the fine-tuned deep learning (DL) models RobBERT and XLM-RoBERTa and all generative models: fine-tuned (FT) and zero-shot (ZS) GEITje, ChocoLlama and Llama prompted in Dutch (NL), as well as Llama prompted in English (EN). Results are provided regarding precision P and recall R, for classes pass (p) and fail (f), and macro-averaged F_1 score.