AI in Education: An Analysis of Large Language Models for Twi Automatic Short Answer Grading

Alex Agyemang and Tim Schlippe(✉)

IU International University of Applied Sciences, Bad Honnef, Germany
tim.schlippe@iu.org

Abstract. Automatic short answer grading can significantly enhance the speed and fairness of grading, making it particularly valuable in areas with a shortage of teachers, such as Africa [1]. However, for most African languages it is very challenging to build automatic short answer grading systems due to the limited availability of natural language processing corpora. Furthermore, only experts can deal with the complex algorithms, required for training and fine-tuning traditional automatic short answer grading systems. Given that state-of-the-art large language models have the potential to address these problems through their growing capabilities and ease of use through prompting, particularly in *zero-shot* and *few-shot* learning, we investigated their performance for grading student answers in the African language Twi. To address the absence of a Twi corpus, we translated and validated the University of North Texas benchmark corpus [2], creating the first Twi automatic short answer grading corpus. On this corpus, we evaluated the performances of the large language models GPT-4o [3], Claude 3 Sonnet [4], and LLaMA 3 [5] as well as for comparison two more traditional approaches: a fine-tuned AfroLM and a cross-lingual M-BERT approach. Among individual models, the cross-lingual M-BERT had the best performance with a mean absolute error of 0.79 points out of 5 points, followed by fine-tuned AfroLM at 0.73 points and Claude 3 Sonnet at 1.00 points. However, combining AfroLM and M-BERT outputs achieved the lowest mean absolute error of 0.64 points, which is less than the human grader variance of 0.75 points in the original corpus [6]. Combining the outputs of the large language models GPT-4o, Claude 3 Sonnet, and LLaMA 3, obtained through *few-shot* learning, yielded a mean absolute error of 1.10 points.

Keywords: AI in Education · Automatic Short Answer Grading · Natural Language Processing · Africa · Large-Language Models · LLMs

1 Introduction

United Nations have issued a global alert about a severe teacher shortage, highlighting the need for 44 million new teachers by 2030 to meet the Sustainable

A. Gerber et al. (Eds.): SACAIR 2024, CCIS 2326, pp. 107–123, 2025.
https://doi.org/10.1007/978-3-031-78255-8_7

Development Goals [1,7]. This shortage is most critical in sub-Saharan Africa, where 15 million teachers are needed [1]. The field of "AI in Education" explores the use and assessment of Artificial Intelligence (AI) techniques within educational and training environments [8–10]. A key focus of this research is to analyze and enhance teaching and learning processes. Advances in automatic short answer grading (ASAG) have progressed to a point where it can support educators by providing faster and fairer grading [6,11].

As the automation associated with ASAG can save money, ASAG is particularly interesting in countries with weaker economies. However, only a few ASAG systems have been developed for languages spoken in these countries. In particular, there is no ASAG corpus or systems for the African language Twi.

Recently, large language models (LLMs) have become widely accessible, and initial analyses have demonstrated their impressive *zero-shot* capabilities across various tasks in English and other languages. However, there has been no research on utilizing LLMs for ASAG in Twi. Therefore, in this study, we evaluate the *zero-shot* and *few-shot* performance of state-of-the-art LLMs GPT-4o [3], Claude 3 Sonnet [4] and LLaMA 3 [5] in automatically grading student answers in Twi. Our contributions are:

- We produce an ASAG corpus for the African language Twi, which we provide to the research community in our GitHub repository[1].
- We analyse state-of-the-art LLMs' ASAG performances for Twi.
- We compare the LLMs' *zero-shot* and *few-shot* performances in Twi ASAG.
- For comparison, we explore fine-tuning AfroLM—a pre-trained African natural language processing model (NLP)—and a cross-lingual M-BERT for ASAG using our new corpus.

In the following section, we will give a brief insight into the linguistic categorization and the peculiarities of Akuapem Twi. In Sect. 3, we will describe related work regarding ASAG, LLMs and NLP for Twi. The experimental setup which includes our analyzed pre-trained language models and LLMs as well as of our ASAG data collection will be presented in Sect. 4. In Sect. 5, we will demonstrate the results of our experiments. Finally, we will summarize our work and indicate possible future steps in Sect. 6.

2 The Language Twi

Twi is a collection of dialects within the Akan language [12]. The number of Twi speakers worldwide is estimated to be around 18 million [13]. The majority of Twi speakers are located in Ghana [14], where it is a widely spoken language among the Akan people. Additionally, there are smaller communities of Twi speakers in neighboring countries such as Cote d'Ivoire and Benin as well as in diaspora communities in countries like Suriname, Jamaica, and the United States [15].

[1] https://github.com/AgyemangOpambour/Twi-Automatic-Short-Answer-Grading/tree/main/Corpus/Twi.

The Akan dialects include Agona, Akuapem, Akwamu, Asante, Akyem, Assin, Bono, Fante, Kwahu, Wassa, Sefwi, Anyi, and Guan [16]. These dialects are classified into two categories: Fante and Twi, with Twi encompassing all non-Fante dialects [16]. Since mutual intelligibility is not universal among the dialects, Akuapem Twi serves as a pivot language for educational purposes in schools. The Akan Orthography Committee (AOC) established a unified Akan orthography in 1978, primarily based on Akuapem Twi [17]. Consequently, Twi dialects without their orthography use Akuapem Twi (marked in red) [18], while Asante and Bono have their writing systems. Fante, not a Twi dialect, has its orthography. Twi is a tonal language with high, mid, and low tones, each carrying distinct semantic meanings [19]. The Twi alphabet comprises 22 letters, including 15 consonants and 7 vowels [20]. Additionally, the letters C, J, V, and Z are used, mostly in loanwords. Twi features 10 diphthongs and many words with multiple meanings that can vary based on context [21]. For instance, "me papa" can mean "good mood" or "my dad," depending on the context.

3 Related Work

In this section, we will describe related work concerning ASAG, LLMs and NLP for Twi.

3.1 Automatic Short Answer Grading

The field of ASAG is increasingly significant as many educational institutions, both public and private, conduct online courses and examinations [6,11]. A comprehensive overview of pre-deep learning ASAG approaches is provided in [22]. Works like [6] and [23] demonstrated that BERT-based deep learning systems outperform others for English and German. [23]'s multilingual RoBERTa model [24] exhibits strong cross-lingual generalization. In [6], ASAG was extended to 26 languages using the smaller M-BERT model [25], achieving mean absolute errors between 0.41 points and 0.72 points out of 5 points, better than the 0.75 points variance between two graders, suggesting AI's viability in supporting answer scoring. Concerning state-of-the-art LLMs, GPT-3.5 and GPT-4 have been evaluated for automatic short answer grading. For instance, [26] examined Finnish student answers in *zero-shot* and *one-shot* settings, finding that GPT-4 significantly outperformed GPT-3.5 in the *one-shot* scenario. Additionally, [27] emphasized the effectiveness of fine-tuned GPT-3.5 for English ASAG. Additionally, [28] indicate that the quantized LLaMA-2 13B model, when fine-tuned with the Short Answer and Feedback dataset [29], exhibits exceptional performance in English ASAG. To the best of our knowledge, no publication describes an analysis of ASAG for African languages.

3.2 Large Language Models

The latest advancements in LLMs have showcased remarkable progress in NLP tasks, driven primarily by innovations in transformer architecture. Models like

GPT-4 [3], GPT-4o, Claude 3 [4], and LLaMA 3 [5] have set new benchmarks in terms of performance and versatility. These models leverage the self-attention mechanisms inherent to transformers, allowing them to process and understand context with unprecedented accuracy [30].

GPT-4, developed by OpenAI, exemplifies this progress with its ability to generate coherent, contextually appropriate text and perform tasks such as summarization, translation, and sentiment analysis with high accuracy [31]. The voice-to-voice capabilities of GPT-4o further extend the model's applicability to speech-related tasks, demonstrating significant advancements in multimodal processing. Claude 3, designed by Anthropic, focuses on alignment and safety, aiming to reduce harmful outputs while maintaining high performance in generative tasks. LLaMA 3, from Meta, offers strong performance in NLP benchmarks, particularly excelling in tasks requiring deep contextual understanding and nuanced language processing. Moreover, in question answering, GPT-4 and Claude 3 have demonstrated remarkable proficiency, achieving high scores on benchmarks such as SQuAD [32].

The success of these LLMs is attributed to their extensive pre-training on diverse and large-scale datasets, enabling them to generalize well across various domains and tasks. As the field continues to evolve, these models set the stage for even more sophisticated and capable NLP systems, promising further advancements in AI. Consequently, we selected GPT-4o, Claude 3 [4], and LLaMA 3 for our experiments.

To the best of our knowledge, there has been no scientific publication that described the use of LLMs for Twi NLP tasks. But other African languages have been analyzed with regards to the use of LLMs. For example, [33] analyzed the performance of four LLMs mT0, Aya, LLaMA 2, and GPT-4 on six tasks: topic classification, sentiment classification, machine translation, summarization, question answering, and named entity recognition across 60 African languages. The evaluation used the LLMs from scratch. The results indicated that all LLMs performed worse on African languages compared to high-resource languages. GPT-4 showed average performance in classification tasks but struggled with generative tasks. mT0 excelled in cross-lingual question answering, Aya was competitive and outperformed mT0 in topic classification, while LLaMA 2 performed the worst, likely due to its English-centric pre-training. The study concludes that there is a significant performance gap for African languages in current LLMs, highlighting the need for improvements, especially in generative tasks and overall support for African languages. [34] investigated the accuracy of LLMs, specifically GPT-4.0, GPT-3.5, and Bard (LaMDA), in answering basic day-to-day financial questions. The study focused on two languages: English and Yoruba, using a financial questions dataset in both languages. The results showed that GPT-4.0 outperformed GPT-3.5 and Bard (LaMDA) in all tested phases. The study suggests that while LLMs can be effective for financial queries, there is potential for improving these models to better handle low-resource languages like Yorùbá.

3.3 Natural Language Processing for Twi

[35] introduced the first transformer-based language models for Twi, specifically focusing on the Akuapem and Asante dialects. Their models, ABENA and BAKO, were fine-tuned on Akan corpora and demonstrated applicability in named entity recognition, neural machine translation, sentiment analysis, and part-of-speech tagging.

Machine translation for Twi has seen notable progress with several key studies. [36] created a transformer-based model using an English-Twi parallel corpus, highlighting the importance of native speakers in refining translations. [37] introduced the Twi-2-ENG parallel corpus from various sources, emphasizing the need for linguistic professionals to enhance corpus quality. [38] developed a Twi-French parallel corpus, showing improved translation performance using English as a pivot. Additionally, [39] addressed gender bias in machine translation, finding that participatory data design can improve translation equality between gendered terms.

Named entity recognition for Twi has been explored through the use of word embeddings and multilingual models. [40] compared word embeddings for Yorùbá and Twi using the Global Voices corpus, demonstrating the effectiveness of multilingual BERT for named entity recognition tasks and highlighting the significance of data quality.

Sentiment analysis for Twi has been advanced using ensemble learning approaches. [41] participated in SemEval 2023's AfriSenti task, focusing on sentiment analysis for multiple African languages, including Twi. They combined multilingual LLMs with language-independent features, achieving notable F1-scores for Twi.

Significant strides have been made in developing automatic speech recognition and synthesis systems for Twi. [42] developed an ASR-based mobile application, TwiGrad, to facilitate learning and maintaining the Twi language. [43] created an end-to-end text-to-speech system for Twi, demonstrating the feasibility of deploying such systems on resource-constrained devices.

Various corpora have been developed to support Twi NLP tasks. The Twi dictionary corpus by [44] is a notable example, consisting of 1,367 words derived from the TypeCraft corpus of Interlinear Glossed Texts. [45] developed a parallel Bible corpus for Twi and other languages, which included aligned sentences and 2D vector representations of Twi vocabulary. Additionally, the Twi-English Parallel Corpus by [19] consists of 25,421 sentence pairs, verified and corrected by native speakers, enhancing the quality of the dataset for machine translation. Furthermore, [38] developed another Twi-French parallel corpus with 10,708 parallel sentences, further contributing to the resources available for Twi language processing. However, to the best of our knowledge, no corpus for Twi ASAG was collected in related work.

4 Experimental Setup

In this section, we will describe our analyzed pre-trained language models and LLMs as well as of our ASAG data collection.

4.1 Overview of the Automatic Short Answer Grading Approaches

An overview of the investigated ASAG approaches for Twi is given in Fig. 1. For our three state-of-the-art LLMs GPT-4o, Claude 3 Sonnet and LLaMA 3, we compare their *zero-shot* and *few-shot* performances. In the *zero-shot* approaches, we instructed the LLMs to assign points between 0 and 5 without giving them examples of exam question, model answer, student answer and the corresponding score assigned by the graders. In contrast, in the *few-shot* approaches, we instructed the LLMs to assign points between 0 and 5 giving them 100 examples of exam question, model answer, student answer and the corresponding score from our Twi ASAG training data.

For comparison, we explore fine-tuning AfroLM—a pre-trained African NLP model—and a cross-lingual M-BERT for ASAG using our Twi ASAG training data. To solve the problems of low-resource languages in building NLP systems, some researchers propose cross-lingual NLP approaches. Thus is it possible to benefit from rich-resource languages like English [46–49]. They usually translate the text from the original low-resource language to English. This allows to do the machine learning task with well-performing models trained with a lot of English resources. Therefore, for the cross-lingual experiment we first machine-translated the Twi text to English and then used the English M-BERT for the grading, which was originally pre-trained with the BooksCorpus (800M words) [50] and English Wikipedia (2,500M words) [51] and then fine-tuned by us using the English ASAG training data.

4.2 Our Analyzed Large Language Models

To compare *zero-shot* and *few-shot* ASAG performances of the state-of-the-art LLMs, we analyzed GPT-4o, Claude 3 Sonnet and LLaMA 3.

GPT-4o. GPT-4o (GPT-4 Omni)[2], developed by OpenAI, was released in May 2024. Unlike GPT-3.5 and GPT-4, which rely on other models to process sound, GPT-4o natively supports voice-to-voice since the model was trained end-to-end across text, vision, and audio, so that all inputs and outputs are processed by the same neural network. While specific training data details for GPT-4o are not publicly disclosed, it can be assumed that it also follows the trend of massive data utilization seen in GPT-3 and GPT-4. For GPT-3, it is known that 570 GB of text data was used [52]. The multilingual and multimodal GPT-4o can process up to 128k tokens per input[3] and supports more than 50 languages[4]. Although there is no specific publication where GPT-4o has been used for ASAG, GPT-3.5 and GPT-4 have been tested to automatically grade student answers. For example, [26] tested Finnish student answers under *zero-shot* and *one-shot* settings. They report a good performance for GPT-4 in the *one-shot* setting, clearly outperforming GPT-3.5. Furthermore, [27] highlight the potential of fine-tuned GPT-3.5 for English ASAG.

[2] https://openai.com/index/hello-gpt-4o.
[3] https://platform.openai.com/docs/models/gpt-4o.
[4] https://openai.com/index/gpt-4o-and-more-tools-to-chatgpt-free.

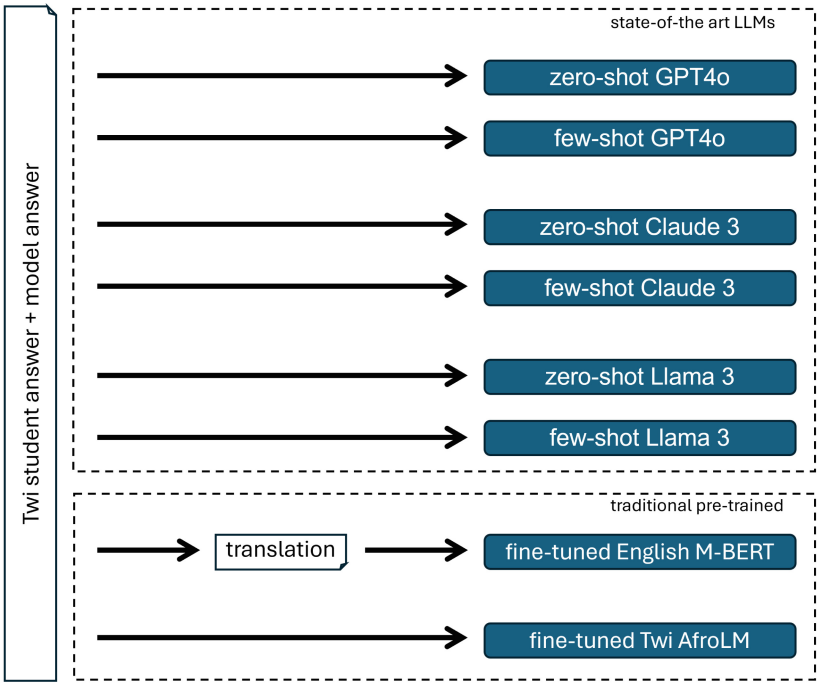

Fig. 1. Overview of the investigated models

Claude 3 Sonnet. Claude 3 Sonnet, developed by Anthropic, was released in late 2023. While specific training data details for the Claude 3 are not publicly disclosed, [4] report that "Claude 3 models are trained on a proprietary mix of publicly available information on the Internet as of August 2023, as well as non-public data from third parties, data provided by data labeling services and paid contractors, and data" that they "generate internally". Claude 3 Sonnet can process up to 200k tokens per input [4]. There are no specific publications detailing the use of Claude 3 Sonnet for ASAG.

LLaMA 3. LLaMA 3 is a large language model provided by Meta AI, launched in early 2024[5]. The model is pre-trained on over 15T tokens that were all collected from publicly available sources[6]. It can process up to 8k tokens per input. There are two LLaMA 3 versions: One model with 8B and one with 70B parameters. We used *LLaMA-3-70B-Instruct*. While there are no specific publications on the use of LLaMA 3 for ASAG, [28] report that the predecessor—quantized LLaMA-2 13B model, fine-tuned with data from the Short Answer and Feedback dataset [29]—shows exceptional English ASAG performance.

[5] https://github.com/meta-LLaMA/LLaMA3.
[6] https://ai.meta.com/blog/meta-LLaMA-3.

4.3 Our Analyzed Pre-trained Language Models

To compare the state-of-the-art LLMs with more traditional approaches of fine-tuning pre-trained language models, we analyzed Afro-LM and a cross-lingual approach using M-BERT.

Afro-LM. AfroLM, developed by [53] in 2022, is a multilingual language model specifically tailored for African languages including Twi. The model was pre-trained on a 0.73 GB dataset of 23 African languages from the news domain, which covers many topics such as health, politics, society, sport, environment [53]. Even the model was pre-trained on only 1.6 MB of Twi text, it outperforms other models like AfriBERTa-Large [53]. AfroLM is able to process 256 tokens in a single input sequence[7]. There are no specific publications detailing its use in ASAG yet. To build an ASAG system for Twi, we fine-tuned AfroLM using the validation set of our Twi ASAG corpus.

M-BERT. Multilingual BERT (Multilingual Bidirectional Encoder Representations from Transformers, M-BERT)[8] is a large language model developed by Google AI and released in 2019 [25,51]. The model was trained on Wikipedia (2,500M words) and BookCorpus [50] (800M words) across 104 languages. M-BERT can process up to 512 tokens per input, offering versatile performance across 104 languages [51]. Although primarily designed for cross-lingual understanding, M-BERT and its related BERT models have been adapted for various tasks, including ASAG, e.g. [54–56]. While most research focused only on one language, [6] investigated ASAG for 26 languages and report that their fine-tuned models are able to fairly score free-text answers in those languages. Since M-BERT was not pre-trained on Twi text, we used the model to build an English ASAG system that is an essential component of our cross-lingual ASAG system.

4.4 Data

To address the absence of a Twi corpus, we translated and validated the University of North Texas benchmark corpus [2], creating the first Twi ASAG corpus. This corpus was used for fine-tuning Afro-LM and for *few-shot* learning of the analyzed LLMs GPT-4o, Claude 3 Sonnet and LLaMA 3 as well as for testing the Twi ASAG performance. The original English corpus was used for training and fine-tuning the English BERT system which is a part of our cross-lingual Twi ASAG system.

Corpus of the University of North Texas. We used the English short answer grading dataset of the University of North Texas [2] to create a Twi ASAG corpus. The advantage of this dataset is that it contains scored student answers, while the answers of other short answer grading corpora, e.g., the SemEval-2013Task7 data

[7] https://github.com/bonaventuredossou/MLM_AL.
[8] https://huggingface.co/google-bert/bert-base-multilingual-cased.

sets [57] are only categorized into 3 classes—there is no point-based grading. [58], [11] and [6] investigated and compared state-of-the-art deep learning techniques for ASAG using the data set. The dataset contains 87 questions with corresponding model answer and on average 28.1 manually graded answers per question about the topic *Data Structures* from under-graduate studies. Each student answer received a score from 0–5 points from two independent graders. We used the average of these two scores as our prediction target and randomly selected 70% of the English ASAG data set (1.818 student answers) for training and 10% (1.82 student answers) for fine-tuning the English BERT system which is a part of our cross-lingual Twi ASAG system. The Twi translation of the remaining 20% (455 student) was used for evaluation.

Creating a Twi Automatic Short Answer Grading Corpus Based on the Corpus of the University of North Texas. In order to create a Twi ASAG corpus based on the Corpus of the University of North Texas [2] for the adaptation, evaluation and comparison of Twi ASAG, we translated all English exam questions, model answers and student answers into Twi using (1) Google's Cloud Translation API[9], which is based on [59], and (2) had the translations corrected by two native speakers who are linguistic experts from the Language Department at Aperade Senior High Technical School in Ghana.

Prompt:

> Please grade the student answer, given the question and the model answer below. You can assign points between 0 and 5, where a completely correct answer received 5 points and a completely false answer receives 0 points. Decimal places are also possible. Please provide the question, model answer, student answer, and the number of points in a table format.
>
> <test set in csv format>

Fig. 2. Prompt for *zero-shot* approaches

4.5 Prompts

Figure 2 and Fig. 3 demonstrate the prompts which we elaborated to have the LLMs do the *zero-shot* and *few-shot* grading.

All prompts were executed in the user interface, not via API, as this was cheaper and the LLMs were able to include the context with the examples from the *prompt for teaching how to grade* for grading in the *few-shot* learning as shown in Fig. 3. For prompt engineering, it was important for us to pass a precise description to the LLMs. We figured out that all LLMs can handle tables in

[9] https://cloud.google.com/translate/docs/reference/rest.

Prompt for teaching how to grade:

> Here are examples of exam questions (question), model answers (desired_answer), student answers (student_answer), and corresponding grades (score_avg) in each line in csv format:
>
> <examples from training set in csv format>
>
> From these examples, please learn how to grade, i.e. assign scores for the student answers. The reason is that in the next prompt you will be provided with exam questions (question), model answers (desired_answer), student answers (student_answer) in csv format, and you need to include the corresponding grades (score_avg) in each line.

Prompt for instructing to grade:

> Thanks for learning how to grade based on the examples provided. Now you will be provided with new exam questions (question), model answers (desired_answer), student answers (student_answer) in csv format in Twi, and you need to include the corresponding grades (score_avg) in each line. Grade all the questions provided within the range of 0 to 5. you can assign decimal values. Display your output in table format with the headings "question", "desired_answer", "student_answer", and "score_avg".
>
> <test set in csv format>

Fig. 3. Prompt for *few-shot* approaches

csv format well. Therefore, we handed over the 100 examples for the *few-shot* learning and the student answers to be graded together with the model answers and the exam questions to the LLMs in csv format. We transferred the csv text with the prompt.

5 Experiments and Results

In this section, we will present the Twi ASAG performances of the more traditional approaches of fine-tuning the pre-trained language models AfroLM and M-BERT. Then, we will explore the *zero-shot* and *few-shot* performances of the state-of-the-art LLMs GPT-4o, Claude 3 Sonnet, and LLaMA 3.

5.1 Pre-trained Language Models

To fine-tune the pre-trained Afro-LM for Twi ASAG, we trained 6 epochs—i.e. the model learned 6 times from the entire training set—with a batch size of 8—i.e. in one training iteration 8 samples were used to update the model

weights—using the AdamW optimizer with an initial learning rate of 0.00002. To fine-tune M-BERT—which was used to grade the Twi student answers after they were machine-translated to English—we trained 6 epochs with a batch size of 16 using the AdamW optimizer with an initial learning rate of 0.00002. We supplemented each model with a linear regression layer that outputs a prediction of the score given an answer. The model expects the model answer and the student answer as input.

Table 1 shows the ASAG performances of the single models, represented by the mean absolute error (MAE). AfroLM, fine-tuned with the Twi ASAG data (*AfroLM*) exhibits the best performance in grading our Twi test set with an MER of 0.73 points out of 5 points. The cross-lingual approach using M-BERT trained with English ASAG data for the machine-translated test set (*TW-EN translation + GPT-4o*$_{zero-shot}$) shows a slightly higher MER of 0.79 points. For comparison, [6] report that their best M-BERT model achieves an MER of 0.45 points for English ASAG, while other systems reach up to 0.86 depending on the language.

Table 1. ASAG performances of single models.

Model	Lang$_{Prompt}$	Lang$_{toGrade}$	MER
AfroLM	—	TW	**0.73**
TW-EN translation + M-BERT	—	EN	0.79
GPT-4o$_{zero-shot}$	TW	TW	2.36
TW-EN translation + GPT-4o$_{zero-shot}$	EN	EN	2.51
GPT-4o$_{zero-shot}$	EN	TW	2.15
GPT-4o$_{few-shot}$	EN	TW	1.53
Claude 3.5 Sonnet$_{zero-shot}$	EN	TW	1.01
Claude 3.5 Sonnet$_{few-shot}$	EN	TW	1.00
LLaMa 3$_{zero-shot}$	EN	TW	1.48
LLaMa 3$_{few-shot}$	EN	TW	1.11

5.2 Large Language Models

Next, we wanted to find out whether the LLM performance is higher when the LLMs are prompted in Twi ($Lang_{Prompt}$ = TW) or in English ($Lang_{Prompt}$ = EN) and whether the performance is higher when the text to be graded is in Twi ($Lang_{toGrade}$ = TW) or in English ($Lang_{toGrade}$ = EN). We analyzed this using GPT-4o with *zero-shot* learning (GPT-4o$_{zero-shot}$). Table 1 indicates that using an English prompt ($Lang_{Prompt}$ = EN) and the original Twi text of the test set to be graded ($Lang_{toGrade}$ = TW) (MER = 2.15 points) performs better than using a Twi prompt ($Lang_{Prompt}$ = TW) (MER = 2.36 points) or grading the test set after is was machine-translated into English ($Lang_{toGrade}$ = EN) (MER = 2.51 points).

5.3 Impact of Providing Examples to Large Language Models

Comparing the *zero-shot* performances—where we passed no ASAG examples to the LLM in the prompt—with the *few-shot* performances—where we passed 100 ASAG examples in the prompt—we see the following: Of all three LLMs, the *zero-shot* and *few-shot* performances of Claude 3 Sonnet are the best, although there is no real difference (1% relative) between *zero-shot* (MER = 1.01 points) and *few-shot* (MER = 1.00 points). In contrast, GPT-4o and LLaMa 3 benefit from the 100 ASAG examples: GPT-$4o_{few-shot}$ (MER = 1.53 points) is 40.52% relatively higher than GPT-$4o_{zero-shot}$ (MER = 2.15 points). LLaMa $3_{few-shot}$ (MER = 1.11 points) outperforms LLaMa $3_{zero-shot}$ (MER = 1.48 points) by 33.33% relative.

5.4 Combination of the Outputs

To further improve ASAG performance, we analyzed a combination of the LLM outputs using the average number of points assigned for each student answer. Our idea was to mitigate the incorrect grading of individual LLMs using this procedure. Table 2 shows that combining AfroLM and M-BERT outputs achieves the lowest MER of 0.64 points, which is 17.19% relatively less than the original corpus's human grader variance of 0.75 points [6]. Combining the outputs of the LLMs GPT-4o, Claude 3 Sonnet, and LLaMA 3, obtained through *zero-shot* learning, yielded an MER of 1.26 points. Combining the outputs of the LLMs GPT-4o, Claude 3 Sonnet, and LLaMA 3, obtained through *few-shot* learning, yielded an MER of 1.10 points, which means a relative improvement of 14.54% compared to *zero-shot*. Combining the outputs of AfroLM and M-BERT and GPT-$4o_{few-shot}$, Claude 3 Sonnet$_{few-shot}$, and LLaMA $3_{few-shot}$ could not mitigate the worse performances of the LLMs and resulted in an MER of 0.99 points as well as combining the outputs of all tested models which resulted in an MER of 1.15 points.

Table 2. ASAG performances of single models.

Model Combination	MER
Pre-trained LMs (AfroLM + M-BERT)	**0.64**
$LLMs_{zero-shot}$ (GPT-4o + Claude 3.5 Sonnet + LLaMa 3)	1.26
$LLMs_{few-shot}$ (GPT-4o + Claude 3.5 Sonnet + LLaMa 3)	1.10
Pre-trained LMs + $LLMs_{few-shot}$	0.99
Pre-trained LMs + $LLMs_{few-shot}$ + $LLMs_{zero-shot}$	1.15

6 Conclusion and Future Work

We have evaluated the performance of the state-of-the-art LLMs GPT-4o, Claude 3 Sonnet, and LLaMA 3 for ASAG in the African language Twi. Due to the

lack of a Twi corpus, we translated and validated the University of North Texas benchmark corpus to create the first Twi ASAG dataset. Our findings reveal that the traditional pre-trained models AfroLM and M-BERT, particularly when their outputs are combined, outperform individual LLMs in terms of MAE. Specifically, combining AfroLM and M-BERT achieved an MAE of 0.64 points, which is lower than the human grader variance. However, training and fine-tuning such systems can only be done by experts due to these tasks' complexity. In contrast, LLMs which are easier to instruct through prompting demonstrate potential even if their performance varies significantly depending on the language of the prompts and the text to be graded, with English prompts generally yielding better results. Additionally, providing LLMs with a few examples improves their performance, though the extent varies across LLMs.

For future work, it would be interesting to investigate whether a weighted combination of LLM outputs leads to improvements. Since LLMs are already good at other tasks without fine-tuning or *few-shot* learning compared to traditional pre-trained models, it could be investigated how well the LLMs explain the points assigned for the student answers. Preliminary work in the area of Explainable AI in ASAG already exists, e.g. [60], but not yet with LLMs and not in an African language such as Twi. Furthermore, it would be interesting to collect a real Twi ASAG corpus and analyze the Twi ASAG performance in exams in other subjects.

References

1. United Nations News: UN Issues Global Alert over Teacher Shortage. UN News (2024). https://news.un.org/en/story/2024/02/1147067
2. Mohler, M., Bunescu, R., Mihalcea, R.: Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. In: The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, Oregon, USA, pp. 752–762. ACL (2011). https://aclanthology.org/P11-1076
3. OpenAI, Achiam, J., Adler, S., Agarwal, S., et al.: GPT-4 Technical Report (2024). https://arxiv.org/abs/2303.08774
4. Anthropic: The Claude 3 Model Family: Opus, Sonnet, Haiku (2024). https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf
5. Touvron, H., et al.: LLaMA: Open and Efficient Foundation Language Models (2023). https://arxiv.org/abs/2302.13971
6. Schlippe, T., Sawatzki, J.: Cross-lingual automatic short answer grading. In: The 2nd International Conference on Artificial Intelligence in Education Technology (AIET), Wuhan, China (2021)
7. United Nations: Sustainable Development Goals: 17 Goals to Transform our World (2024). https://www.un.org/sustainabledevelopment/sustainabledevelopment-goals. Accessed Aug 2024
8. Chen, L., Chen, P., Lin, Z.: Artificial intelligence in education: a review. IEEE Access **8**, 75264–75278 (2020). https://doi.org/10.1109/ACCESS.2020.2988510

9. Schlippe, T., Cheng, E.C.K., Wang, T.: Artificial Intelligence in Education Technologies: New Development and Innovative Practices. Springer, Cham (2023). https://doi.org/10.1007/978-981-99-7947-9
10. Cheng, E.C.K., Wang, T., Schlippe, T., Beligiannis, G.N.: Artificial Intelligence in Education Technologies: New Development and Innovative Practices. Springer, Cham (2023). https://doi.org/10.1007/978-981-19-8040-4
11. Sawatzki, J., Schlippe, T., Benner-Wickner, M.: Deep learning techniques for automatic short answer grading: predicting scores for English and German answers. In: Cheng, E.C.K., Koul, R.B., Wang, T., Yu, X. (eds.) Artificial Intelligence in Education: Emerging Technologies, Models and Applications. LNDECT, vol. 104, pp. 65–75. Springer, Singapore (2022). https://doi.org/10.1007/978-981-16-7527-0_5
12. Akan (Twi) at Rutgers (2022). https://www.amesall.rutgers.edu/languages/128-akan-twi. Accessed Jan 2023
13. Yakubu, M.: Check Out Other Countries That Speak Twi Apart From Ghana (2024). https://www.primenewsghana.com/entertainment/check-out-other-countries-that-speak-twi-apart-from-ghana.html. Accessed 29 July 2024
14. Akan Twi (2023). https://celt.indiana.edu/portal/Akan%20Twi/index.html. Accessed Jan 2023
15. Lingual, P.: Check Out 4 Countries Where They Speak Twi Aside Ghana That You Never Knew (2024). https://paullingual.com/check-out-4-countries-where-they-speak-twi-aside-ghana-that-you-never-knew-check-out. Accessed 29 July 2024
16. Osam, E.K.: An introduction to the verbal and multi-verbal system of akan. In: Workshop on Multi-verb Constructions, Trondheim, Norway (2003)
17. Kouadio, N.J.: A Unified Orthography for the Akan Languages of Ghana and Ivory Coast: General Unified Spelling Rules, Monograph Series/Centre for Advanced Studies of African Society, vol. 20. Centre for Advanced Studies of African Society, CASAS, Cape Town (2003)
18. Schachter, P., Fromkin, V.: A Phonology of Akan: Akuapem, Asante, Fante. Working papers in phonetics, University of California (1979)
19. Azunre, P., et al.: English-Twi Parallel Corpus for Machine Translation. arXiv abs/2103.15625 (2021)
20. The African Linguists Network Blog: Language Guide. https://alnresources.wordpress.com/african-culture-and-language. Accessed 15 May 2023
21. Alabi, J.O., Amponsah-Kaakyire, K., Adelani, D.I., España-Bonet, C.: Massive vs. curated embeddings for low-resourced languages: the case of Yorùbá and Twi. In: The 12th Conference on Language Resources and Evaluation (LREC 2020) (2020)
22. Burrows, S., Gurevych, I., Stein, B.: The eras and trends of automatic short answer grading. Int. J. Artif. Intell. Educ. **25**(1), 60–117 (2015). https://doi.org/10.1007/s40593-014-0026-8
23. Camus, L., Filighera, A.: Investigating transformers for automatic short answer grading. In: Bittencourt, I.I., Cukurova, M., Muldner, K., Luckin, R., Millán, E. (eds.) AIED 2020. LNCS (LNAI), vol. 12164, pp. 43–48. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-52240-7_8
24. Liu, Y., et al.: RoBERTa: A Robustly Optimized BERT Pretraining Approach. CoRR abs/1907.11692 (2019). http://arxiv.org/abs/1907.11692
25. Pires, T., Schlinger, E., Garrette, D.: How multilingual is multilingual BERT? In: The 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, pp. 4996–5001. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/P19-1493. https://aclanthology.org/P19-1493

26. Chang, L.H., Ginter, F.: Automatic short answer grading for finnish with ChatGPT. In: The AAAI Conference on Artificial Intelligence, vol. 38, no. 21, pp. 23173–23181 (2024). https://doi.org/10.1609/aaai.v38i21.30363. https://ojs.aaai.org/index.php/AAAI/article/view/30363
27. Latif, E., Zhai, X.: Fine-tuning ChatGPT for automatic scoring. Comput. Educ. Artif. Intell. **6**, 100210 (2024). https://doi.org/10.1016/j.caeai.2024.100210. https://www.sciencedirect.com/science/article/pii/S2666920X24000110
28. Katuka, G.A., Gain, A., Yu, Y.Y.: Investigating Automatic Scoring and Feedback using Large Language Models (2024). https://arxiv.org/abs/2405.00602
29. Filighera, A., Parihar, S., Steuer, T., Meuser, T., Ochs, S.: Your answer is incorrect... would you like to know why? Introducing a bilingual short answer feedback dataset. In: The 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Dublin, Ireland, pp. 8577–8591. Association for Computational Linguistics (2022). https://doi.org/10.18653/v1/2022.acl-long.587. https://aclanthology.org/2022.acl-long.587
30. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30. Curran Associates, Inc. (2017)
31. Brown, T., et al.: Language models are few-shot learners. In: Advances in Neural Information Processing Systems, vol. 33, pp. 1877–1901. Curran Associates, Inc. (2020)
32. Raffel, C., et al.: Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. **21**(140), 1–67 (2020). http://jmlr.org/papers/v21/20-074.html
33. Ojo, J., Ogueji, K., Stenetorp, P., Adelani, D.I.: How good are Large Language Models on African Languages? (2024). https://arxiv.org/abs/2311.07978
34. Sikiru, R.D., Adekanmbi, O., Soronnadi, A.: Comparative study of LLMs for personal financial decision in low resource language. In: 5th Workshop on African Natural Language Processing (2024). https://openreview.net/forum?id=9gDt0ZUk8H
35. Azunre, P., et al.: Contextual text embeddings for Twi. In: 2nd AfricaNLP Workshop Proceedings, AfricaNLP@EACL 2021, Virtual Event, 19 April 2021 (2021). https://arxiv.org/abs/2103.15963
36. Bannerman, S., Agyei, E., Sarpong, S., Quaye, A.B., Yussif, S.B., Agbesi, V.K.: Machine translation from English-Twi in parallel corpus: low resource Ghanaian. Language (2023). https://doi.org/10.2139/ssrn.4761197
37. Agyei, E., Zhang, X., Bannerman, S., et al.: Low resource Twi-English parallel corpus for machine translation in multiple domains (Twi-2-ENG). Discov. Comput. **27**, 17 (2024). https://doi.org/10.1007/s10791-024-09451-8
38. Gyasi, F., Schlippe, T.: Twi machine translation. Big Data Cogn. Comput. **7**(2) (2023). https://doi.org/10.3390/bdcc7020114. https://www.mdpi.com/2504-2289/7/2/114
39. Oppong, A.: Building a participatory data design approach to examine gender bias in english-twi machine translation. In: Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems. CHI EA 2023. Association for Computing Machinery, New York (2023). https://doi.org/10.1145/3544549.3583942
40. Alabi, J.O., Amponsah-Kaakyire, K., Adelani, D.I., España-Bonet, C.: Massive vs. curated embeddings for low-resourced languages: the case of Yorùbá and Twi. In: International Conference on Language Resources and Evaluation (2019)
41. García-Díaz, J.A., Caparros-laiz, C., Almela, Á., Alcaráz-Mármol, G., Marín-Pérez, M.J., Valencia-García, R.: UMUTeam at SemEval-2023 task 12: ensemble learning of LLMs applied to sentiment analysis for low-resource African lan-

guages. In: The 17th International Workshop on Semantic Evaluation (SemEval-2023), Toronto, Canada, pp. 285–292. Association for Computational Linguistics (2023). https://doi.org/10.18653/v1/2023.semeval-1.38. https://aclanthology.org/2023.semeval-1.38
42. Quartson, P.: TWIGRAD: An ASR-Based Application for Learning Twi, applied Project, Department of Computer Science, Ashesi University College. B.Sc, Computer Science (2021)
43. Aboagye, F., Akolly, E.: Text-to-Speech for Ghanaian Language (Akuapem Twi) on an Embedded System, capstone Project, Department of Engineering, Ashesi University College. B.Sc, Electrical/Computer Engineering (2021)
44. Beermann, D., Hellan, L., Mihaylov, P., Struck, A.: Developing a Twi (Asante) dictionary from Akan interlinear glossed texts. In: The 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL), Marseille, France, pp. 294–297. ELRA (2020). https://aclanthology.org/2020.sltu-1.41
45. Adjeisah, M., Liua, G., Nortey, R.N., Song, J.: English Twi parallel-aligned bible corpus for encoder-decoder based machine translation. Acad. J. Sci. Res. **8**(12), 371–382 (2020)
46. Balahur, A., Turchi, M.: Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. Comput. Speech Lang. **28**, 56–75 (2014)
47. Lin, Z., Jin, X., Xu, X., Wang, Y., Tan, S., Cheng, X.: Make it possible: multilingual sentiment analysis without much prior knowledge. In: IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), vol. 2, pp. 79–86 (2014). https://doi.org/10.1109/WI-IAT.2014.83
48. Vilares, D., Alonso Pardo, M., Gómez-Rodríguez, C.: Supervised sentiment analysis in multilingual environments. Inf. Process. Manag. **53** (2017). https://doi.org/10.1016/j.ipm.2017.01.004
49. Can, E.F., Ezen-Can, A., Can, F.: Multilingual sentiment analysis: an RNN-based framework for limited data. In: ACM SIGIR 2018 Workshop on Learning from Limited or Noisy Data (2018)
50. Zhu, Y., et al.: Aligning books and movies: towards story-like visual explanations by watching movies and reading books. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 19–27 (2015). https://doi.org/10.1109/ICCV.2015.11
51. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: The 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, Minnesota, pp. 4171–4186. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/N19-1423. https://aclanthology.org/N19-1423
52. Brown, T.B., et al.: Language models are few-shot learners. In: The 34th International Conference on Neural Information Processing Systems. NIPS 2020. Curran Associates Inc., Red Hook (2020)
53. Dossou, B.F.P., et al.: AfroLM: a self-active learning-based multilingual pretrained language model for 23 African languages. In: The Third Workshop on Simple and Efficient Natural Language Processing (SustaiNLP), Abu Dhabi, United Arab Emirates (Hybrid), pp. 52–64. Association for Computational Linguistics (2022). https://doi.org/10.18653/v1/2022.sustainlp-1.11. https://aclanthology.org/2022.sustainlp-1.11

54. Sung, C., Dhamecha, T., Saha, S., Ma, T., Reddy, V., Arora, R.: Pre-training BERT on domain resources for short answer grading. In: The 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, pp. 6071–6075. Association for Computational Linguistics (2019). https://doi.org/10.18653/v1/D19-1628. https://aclanthology.org/D19-1628
55. Ghavidel, H.A., Zouaq, A., Desmarais, M.C.: Using BERT and XLNET for the automatic short answer grading task. In: The 12th International Conference on Computer Supported Education - Volume 1: CSEDU, pp. 58–67. INSTICC, SciTePress (2020). https://doi.org/10.5220/0009422400580067
56. Zhu, X., Wu, H., Zhang, L.: Automatic short-answer grading via BERT-based deep neural networks. IEEE Trans. Learn. Technol. **15**(3), 364–375 (2022). https://doi.org/10.1109/TLT.2022.3175537
57. Dzikovska, M., et al.: SemEval-2013 task 7: the joint student response analysis and 8th recognizing textual entailment challenge. In: Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: The Seventh International Workshop on Semantic Evaluation (SemEval 2013), Atlanta, Georgia, USA, pp. 263–274. Association for Computational Linguistics (2013). https://aclanthology.org/S13-2045
58. Gomaa, W.H., Fahmy, A.A.: Ans2vec: a scoring system for short answers. In: Hassanien, A.E., Azar, A.T., Gaber, T., Bhatnagar, R., F. Tolba, M. (eds.) AMLTA 2019. AISC, vol. 921, pp. 586–595. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-14118-9_59
59. Johnson, M., et al.: Google's multilingual neural machine translation system: enabling zero-shot translation. Trans. Assoc. Comput. Linguist. **5**, 339–351 (2017). https://doi.org/10.1162/tacl_a_00065. https://aclanthology.org/Q17-1024
60. Schlippe, T., Stierstorfer, Q., Koppel, M.T., Libbrecht, P.: Explainability in automatic short answer grading. In: Cheng, E.C.K., Wang, T., Schlippe, T., Beligiannis, G.N. (eds.) AIET 2022, pp. 69–87. Springer, Cham (2023). https://doi.org/10.1007/978-981-19-8040-4_5