# LLMs for Automatic Response Evaluation: Comparaison of Pretrained Transfer Learning Models and Ethical Considerations

Yakhou Yousra[a,b], Bellmir Omar[b], Limouri Saad[b], Benmimoune Oussama[b], Chtioui Rime[b], Bichri Imane[b]

[a] *yousra.yakhou@centrale-casablanca.ma*
[b] *Ecole Centrale Casablanca*

**Abstract**

This paper explores the integration of Language Models (LLMs) in Automatic Short Answer Grading (ASAG), focusing on both technical and ethical considerations. From a technical perspective, the study examines model architecture and compares transfer learning models to enhance LLMs efficiency. The study utilizes pretrained embeddings from transfer learning models, namely ELMo, BERT and GPT-2, to gauge their effectiveness compared to previous approaches. By training on a singular feature, namely cosine similarity, derived from these models, the paper assesses RMSE scores and correlation measurements. The findings reveal that ELMo outperforms the other three models specifically on the Mohler dataset. Ethically, the paper scrutinizes concerns such as bias, fairness, and privacy, emphasizing the broader societal impact of automated assessment.

*Keywords:* Automatic Short Answer Grading, BERT, Embeddings, Ethical Considerations, ELMo, GPT-2, Grading Automation, Transfer Learning Models

## 1. Introduction

The dynamic evolution of technology has profoundly reshaped various facets of our society, and the field of education is no exception to this pervasive digital transformation. Amidst the intricate landscape of the educational journey, the assessment of students, particularly through the nuanced practice of paper grading, stands out as a pivotal juncture. Addressing the escalating demands for efficiency and precision in this realm, Language Modeling Systems (LLMs) have emerged as promising tools.

Recent strides in LLMs, exemplified by models such as Embeddings from Language Models (ELMo), Bidirectional Encoder Representations from Transformers (BERT), and Generative Pretraining (GPT-2), have yielded robust transfer learning models adept at handling a spectrum of tasks. These models, pretrained on extensive datasets, showcase the ability to extract semantic context through resilient architectures.[1]

However, as the integration of these LLMs into educational contexts intensifies, a nexus of technical and ethical challenges comes to the forefront. On the technical front, a thorough exploration of the capabilities, limitations, and challenges in deploying these models for automated paper grading becomes imperative. Questions arise: How adeptly can LLMs evaluate the diversity of student responses? What criteria govern the accuracy and reliability of these systems? An in-depth analysis is crucial to ensure a pertinent and equitable application of this transformative technology. Simultaneously, ethical considerations must not be overlooked. A profound exploration of the ethical dimensions entwined with the assimilation of LLMs into the grading process is essential. Pivotal ethical concerns include the confidentiality of student data, transparency in algorithmic decision-making, and the potential biases inherent in automated assessment. These considerations prompt fundamental questions about the ethical responsibilities held by educators and developers and the fundamental rights of students to an assessment process that is not only technologically advanced but also fair and objective.

In this comprehensive exploration, our study incorporates a rigorous comparative analysis of transfer learning models, delineating their strengths and limitations in the context of grading student papers. The process involves extracting semantic knowledge from responses using embeddings from transfer learning models, encoded with contextual vectors. We adopt a two-step preprocessing approach for regression model training, exploring the potential of transfer learning models like ELMo, BERT, and GPT-2. Their effectiveness is evaluated on the Mohler dataset, focusing on training with cosine similarity and assessing RMSE scores and correlation measurements in comparison to previous methods.

The article delves into prior research on Automatic Short Answer Grading (ASAG), followed by an analysis providing details about the dataset. Subsequently, a concise elucidation of transfer learning models and embeddings is presented, outlining the experimentation process. Finally, the article offers the results, observations derived from the results, and suggestions for potential enhance-

ments, approaching these aspects in a distinctly different manner. In summary, our contribution is a comparaison between three models on short answer grading system which, we also release as open-source software at: yousrayk/Comparaison_of_models_for_ASAG.git

## 2. Previous Works

Pérez and al.[14] introduced corpus-based methods to Automated Short Answer Grading (ASAG), utilizing a combination of Latent Semantic Analysis (LSA) and Bilingual Evaluation Understudy (BLEU) scores. Building on this, Mohler and Mihalcea[11] combined multiple knowledge-based and corpus-based features to extract similarity measures between students' and teachers' answers. This initiative paved the way for incorporating corpus-based methods into machine learning systems[10].

Taking a more feature-rich approach, Sultan et al.[19] extracted various features, including word-alignment, vector similarity, and term frequency–inverse document frequency (tf-idf), and then trained and evaluated their model on the Mohler dataset[10]. In another avenue, Metzler[9]harnessed conventional embeddings in Natural Language Processing (NLP), such as Word2Vec [9], GloVe [13], and FastText [6], to extract distributional and semantic features. Notably, Metzler[9] compared results with pretrained and domain-specific trained word embeddings of Word2Vec, GloVe, and FastText on the Mohler dataset, but conventional embeddings overlooked word context and long-term dependencies.

Recent advancements in LLMs have shifted focus towards models like ELMo [15], BERT [4], and GPT-2 [16], which consider word contexts in sentences and are pretrained on extensive corpora. Additionally, previous works often employed multiple features to evaluate answers, while our approach utilizes a single semantic feature extracted from transfer learning models for training on the same dataset.

## 3. Methodology

### *3.1.* ***The Dataset***

We will be using Mohler's Dataset (Mohler et al., ACL 2011)[10]. Students engaged with 80 questions distributed across ten assignments and two exams, with a total of 31 participants actively submitting responses. The dataset curated by Mohler includes 2273 student responses, slightly fewer than the expected count of 2480, as some students opted not to submit answers for specific assignments.

For the assessment of responses, two independent human judges, both graduate students in the computer science department, evaluated answers on a scale ranging from 0 (completely incorrect) to 5 (perfect answer). Grader1, who served as the teaching assistant for the Data Structures class, and grader2, one of the authors, contributed to this evaluation process. The average grade determined by these two annotators serves as the benchmark against which we compare the output of our system.

The dataset exhibits a bias favoring correct answers, as indicated by (Mohler et al., ACL 2011)[10] with an average grade mean of 4.2 and a median of 4.5 . This evident bias is reflected in the associated figure 1 below.

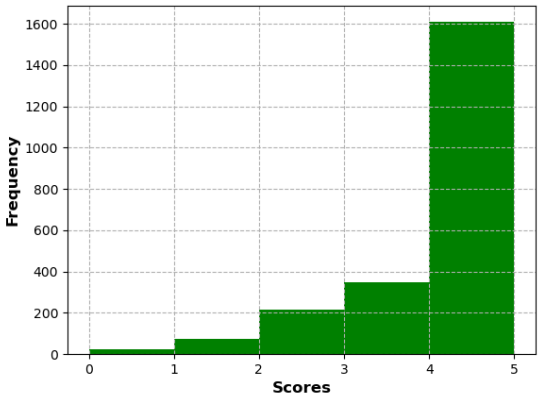


Figure 1: Bar Chart of assigned average scores in Mohler's dataset highlighting the bias towards correct answers.

### *3.2.* ***Transfer Learning Models***

We selected ELMo, BERT, and GPT-2 for short answer grading due to their proven effectiveness in handling diverse linguistic tasks and their robust performance in transfer learning. These models have demonstrated a capacity to capture nuanced semantic contexts and linguistic intricacies, making them well-suited for the nuanced nature of short answer grading. ELMo excels in contextual embeddings[5], BERT is renowned for bidirectional contextualized representations [20], and GPT-2 showcases advanced generative pretraining capabilities. Leveraging the strengths of these models, we aim to enhance the accuracy and efficiency of short answer grading by leveraging their comprehensive understanding of language nuances and contextual information.

#### *3.2.1. Embeddings from Language Models (ELMO)*

ELMo (Embeddings from Language Models) [15] ,is a deep contextualized word representation model that generates word embeddings by capturing the complex contextual meanings of words within sentences. Unlike traditional word embeddings, ELMo considers both the context in which a word appears and its morphological structure by utilizing a bidirectional LSTM (Long Short-Term Memory) trained on a large corpus of text. This allows ELMo to produce embeddings that reflect the nuanced and varying meanings of words in different contexts, enabling more accurate and contextually aware natural language processing tasks.[15]

*Architecture Overview*

ELMo uses vectors derived from a bidirectional LSTM that is trained with a coupled language model (LM) objective on a large text corpus [15]. For this reason, we call them ELMo (Embeddings from Language Models) representations. Unlike previous approaches for learning contextualized word vectors, ELMo representations are deep, in the sense that they are a function of all the internal layers of the biLM. More specifically, we learn a linear combination of the vectors stacked above each input word for each end task, which markedly improves performance over just using the top LSTM layer. As the figure 2

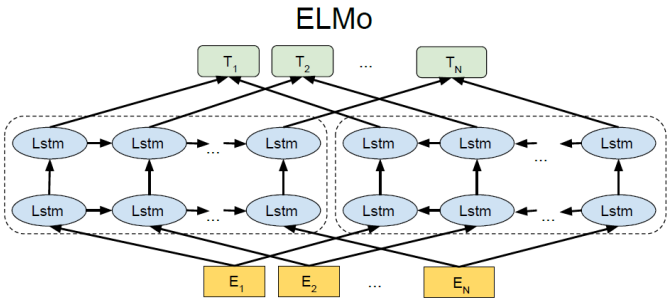


Figure 2: Architecture Overview of ELMO.

*Bidirectional Language Models*

Bidirectional language models biLM are a type of language model that consider both preceding and following words when predicting a word in a sequence.

Unlike most widely used word embeddings, ELMo word representations are functions of the entire input sentence, as described in this section. They are computed on top of two-layer biLMs with character convolutions, as a linear function of the internal network states. This setup allows us to do semi-supervised learning, where the biLM is pretrained at a large scale and easily incorporated into a wide range of existing neural NLP architectures.[8]

Given a sequence of N tokens, a forward language model computes the probability of the sequence by modeling the probability of token $t_k$ given the history:

$$p(t_1, t_2, \ldots, t_N) = \prod_{k=1}^{N} p(t_k | t_1, t_2, \ldots, t_{k-1}) \qquad (1)$$

A backward LM is similar to a forward LM, except it runs over the sequence in reverse, predicting the previous token given the future context:

$$p(t_1, t_2, \ldots, t_N) = \prod_{k=1}^{N} p(t_k | t_{k+1}, t_{k+2}, \ldots, t_N) \qquad (2)$$

At each position $k$, each LSTM layer outputs a context-dependent representation:

$$\vec{\mathbf{h}}_{k,L}^{\mathrm{LM}}$$

When $j = 1, \ldots, L$, the top layer LSTM output $h_{LM}^{k,L}$ is used to predict the next token $t_{k+1}$ with a Softmax layer. It can be implemented in an analogous way to a forward LM, with each backward LSTM layer $j$ in an $L$-layer deep model producing representations of $t_k$ given $(t_{k+1}, t_{k+2}, \ldots, t_N)$. A biLM combines both a forward and backward LM. Our formulation jointly maximizes the log likelihood of the forward and backward directions. [12]

$$\sum_{k=1}^{N} \Bigg[ \log \Big( p(t_k | t_1, t_2, \ldots, t_{k-1}; \{\Theta\}_x, \vec{\Theta}_{LSTM}, \Theta_s) \Big) + \log \Big( p(t_k | t_{k+1}, t_{k+2}, \ldots, t_N; \Theta_x, \overleftarrow{\Theta}_{LSTM}, \Theta_s) \Big) \Bigg] \qquad (3)$$

$\theta_x$ (Token Representation Parameters): These parameters are associated with the token representation or the hidden states learned during the processing of input tokens (words, characters, etc.). In many natural language processing tasks, like language modeling or sequence-to-sequence tasks, $\theta_x$ refers to the weights and biases involved in transforming input tokens into a meaningful hidden representation.

$\theta_s$ (Softmax Layer Parameters): These parameters are specific to the softmax layer, often used in classification tasks. The softmax layer typically comes at the end of the network and is responsible for converting the output of the previous layers into probabilities for different classes. $\theta_s$ includes weights and biases associated with this final layer, where the outputs are transformed into a probability distribution over the classes.

Finally, ELMo is a task-specific combination of the intermediate layer representations in the biLM. For each token $t_k$, an L-layer biLM computes a set of $2L + 1$ representations.

$$R_k = (\mathbf{x}_k^{LM}, \vec{\mathbf{h}}_{k,L}^{LM}, \overleftarrow{\mathbf{h}}_{k,L}^{LM}) \qquad (4)$$

*3.2.2. Bidirectional Encoder Representations from Transformers (BERT)*

BERT stands for Bidirectional Encoder Representations from Transformers [4]. Unlike recent language representation models, BERT is designed to pretrain deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers.[7] As a result, the pre-trained BERT model can be finetuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task specific architecture modifications. BERT is conceptually simple and empirically powerful.[6]

*Architecture Overview*

To make BERT handle a variety of down-stream tasks, the input representation is able to unambiguously represent both a single sentence and a pair of sentences (e.g.

'Question, Answer') in one token sequence [3]. Throughout this work, a "sentence" can be an arbitrary span of contiguous text, rather than an actual linguistic sentence. [4] A "sequence" refers to the input token sequence to BERT, which may be a single sentence or two sentences packed together, as depicted in the figure3 below.

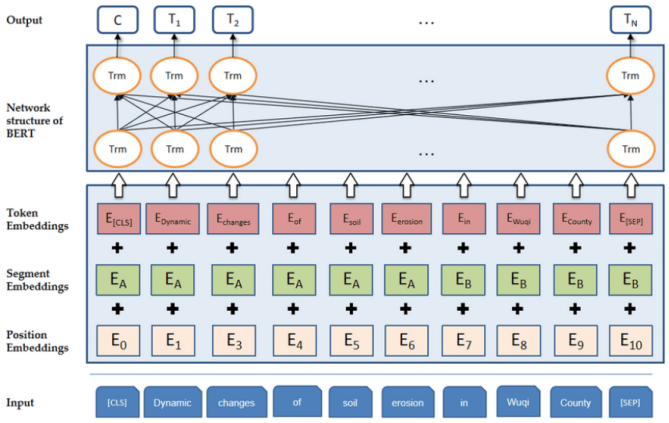


Figure 3: Architecture Overview of BERT.

### *3.2.3. GPT-2*

*GPT-2 Overview:*

GPT-2, or Generative Pre-trained Transformer 2, is a state-of-the-art language model developed by OpenAI. It belongs to the family of transformer-based models and is renowned for its ability to generate coherent and contextually relevant text. GPT-2 is a successor to the original GPT, and one of its distinctive features is its massive scale, boasting a significantly increased number of parameters.

*Architecture of GPT-2:*

GPT-2 inherits its architecture from the transformer model, a groundbreaking neural network architecture introduced by Vaswani et al. in the paper "Attention is All You Need." Here are key aspects of GPT-2's architecture:

1. **Transformer Architecture:** GPT-2 leverages the transformer architecture, which revolves around the concept of self-attention. This architecture has proven highly effective in capturing contextual information from input sequences, making it well-suited for language modeling tasks.

2. **Layered Structure:** The model consists of multiple layers, with each layer containing a multi-head self-attention mechanism and a feedforward neural network. This layered structure enables the model to learn hierarchical representations of input data.

3. **Positional Encoding:** To account for the sequential nature of input sequences, GPT-2 incorporates positional encodings. These encodings help the model understand the position of tokens within a sequence, preserving order information.

4. **Unidirectional Self-Attention:** GPT-2 employs unidirectional self-attention, meaning that each token attends to all preceding tokens in the sequence. This unidirectional approach maintains an autoregressive property during training.

5. **Pretraining and Finetuning:** GPT-2 follows a two-step training process. Initially, it undergoes unsupervised pretraining on a large corpus, enabling it to learn language patterns. Subsequently, the model can be finetuned on specific tasks using labeled data.

6. **WebText Dataset:** GPT-2 is pretrained on the WebText dataset, obtained through web scraping. This dataset is chosen for its diversity, capturing a wide range of language usage from the web.

## *3.3. Details of the experiment*

### *3.3.1. Preprocessing*

In the first step of getting our data ready for analysis, we only broke down sentences into individual words **(tokenization)**.

### *3.3.2. Feature extraction*

a. **Token Assignment:**

Each transfer learning model's pretrained embeddings are linked to the tokens(words or subwords), present in all the answers. This means that every word in each answer is uniquely represented by a pretrained embedding obtained from the respective transfer learning model. This token-to-embedding association allows us to leverage the learned knowledge and semantic understanding embedded in the pretrained embeddings for further analysis and interpretation of the text data.

b. **Sum of Word Embeddings (SOWE)**

The answer embedding is generated using the Sum of Word Embeddings (SOWE) technique. This involves calculating the answer embedding "a" by summing up the pretrained embeddings of all the individual words within that answer. In other words, the answer embedding is derived by adding together the embeddings associated with each word in the response. The outcome is a unified vector that serves as a comprehensive representation of the entire answer, encapsulating the semantic information and contextual meaning conveyed by the combined word embeddings. This approach enables a condensed yet informative representation of each answer for further analysis and model interpretation. The calculation is expressed by the formula:

$$a = \sum_{i=1}^{n} word_i$$

Where $n$ is the number word embeddings in the answer $a$ .

c. **Cosine Similarity:**

In the assessment of student answers, the degree of similarity to the desired answer is evaluated through the application of cosine similarity. Cosine similarity is a mathematical metric that gauges the cosine of the angle formed between two vectors, yielding a numerical value within the range of -1 to 1. A cosine similarity of 1 signifies that the vectors are identical, whereas a value of 0 indicates orthogonality, implying no similarity. Conversely, a cosine similarity of -1 denotes that the vectors are entirely opposite. This method provides a quantitative measure of how closely a student's response aligns with the expected or correct answer, offering a nuanced assessment of the quality and accuracy of the given response. The calculation is expressed by the formula:

$$\cos(\mathbf{a}_{ij}, \mathbf{a}_i) = \frac{\mathbf{a}_{ij} \cdot \mathbf{a}_i}{\|\mathbf{a}_{ij}\| \|\mathbf{a}_i\|}$$

Where each $(a_{ij})$ is $i$ student's answer for question $j$ for student $i$ and $(a_j)$ is the desired answer for question $j$.

d. **Feature Representation**

In feature representation, we use scores as features for answers. These scores measure how much each student's response is similar to the correct answer. It's a straightforward way to compare and evaluate student answers in a standardized way. The preceding studies commonly employ various features to assess responses. In contrast, our approach relies on a singular semantic feature derived from transfer learning models for training. In fact, Sultan et al.[19] conducted an analysis involving multiple features, including word alignment, vector similarity, term frequency–inverse document frequency (tf-idf), and conducted training and evaluation on the Mohler dataset.

#### *3.3.3.* ***Training with Regression Methods***

Regression methods, specifically designed for forecasting continuous outcomes, are utilized for this purpose. To ensure the model's effectiveness and generalizability, the Mohler data is randomly divided into two sets: 70% for training and 30% for testing. During training, the cosine similarity feature is paired with its corresponding grades, and both linear and non-linear (ridge) regression techniques are employed to optimize the model's predictive capabilities. During the evaluation phase, the performance of the trained regression model is assessed using two key metrics: the Root Mean Square Error (RMSE) and the Pearson correlation $\rho$. The RMSE provides a measure of the average deviation between the predicted scores and the actual desired scores, offering insight into the overall accuracy of the model. Simultaneously, the Pearson correlation quantifies the linear relationship between the predicted and desired scores, indicating the strength and direction of the association. These metrics serve as crucial indicators for evaluating the effectiveness and precision of the regression model in capturing and predicting the continuous outcomes of interest.

## 4. Results

The table 1 presents regression metrics, including Pearson correlation ($\rho$) and Root Mean Squared Error (RMSE), for three different models: ELMo, BERT, and GPT-2. In terms of Pearson correlation, ELMo outperforms both BERT and GPT-2 across all regression types—linear, ridge, and isotonic. Specifically, ELMo demonstrates higher correlation coefficients (0.45, 0.449, and 0.482) compared to BERT (0.270, 0.273, and 0.322) and GPT-2 (0.28, 0.270, and 0.270) in linear, ridge, and isotonic regression, respectively. Moreover, ELMo achieves the lowest RMSE values (0.997, 0.997, and 0.980) compared to BERT (1.078, 1.076, and 1.058) and GPT-2 (1.079, 1.078, and 1.600) across the three regression models.

In summary, ELMo consistently demonstrates superior performance in aligning predicted and desired scores, making it the most effective model among the three.

## 5. Discussion

ELMo outperformed GPT-2 and BERT in short answer grading because of its strength in capturing nuanced contextual meanings through bidirectional LSTM. ELMo's focus on both context and morphological structure proved beneficial for understanding the subtleties in student responses. In contrast, BERT's bidirectional contextualized representations and GPT-2's generative pretraining might not be optimized for precise regression tasks like scoring short answers. ELMo's ability to provide a more accurate representation of language nuances likely contributed to its consistent superior performance in aligning predicted scores with actual scores.

## 6. Comparative Analysis with Previous Work on the Mohler Dataset

Notably, the table 2 highlights that, despite ELMo outperforming BERT and GPT-2 in terms of Root Mean Square Error (RMSE) and Pearson Correlation, an interesting observation emerges. The traditional tf-idf approach presented by Sultan et al. in 2016 consistently outperforms all the considered models. This unexpected result underscores the importance of revisiting and acknowledging the effectiveness of established methodologies, even in the face of advancements in more recent and sophisticated techniques. It prompts a deeper reflection on the nuances of the Mohler dataset and the potential strengths inherent in traditional approaches that may warrant further investigation.

| Model | ELMo | | BERT | | GPT-2 | |
|---|---|---|---|---|---|---|
| **Metrics** | $\rho$ | **RMSE** | $\rho$ | **RMSE** | $\rho$ | **RMSE** |
| **Linear Regression** | 0.45 | 0.997 | 0.270 | 1.078 | 0.28 | 1.079 |
| **Ridge Regression** | 0.449 | 0.997 | 0.273 | 1.076 | 0.270 | 1.078 |
| **Istonic Regression** | **0.482** | **0.980** | 0.322 | 1.058 | 0.270 | 1.600 |

Table 1: Regression Metrics for ELMo, BERT, and GPT-2

| Model/Approach | Features | RMSE | Pearson Correlation |
|---|---|---|---|
| BOW (Mohler et al., 2011) | SVMRank | 1.042 | 0.480 |
| | SVR | 0.999 | 0.431 |
| tf-idf (Mohler et al., 2011) | SVR | 1.022 | 0.327 |
| tf-idf (Sultan et al., 2016) | LR + SIM | **0.887** | **0.592** |
| Word2Vec (Metzler, 2019) | SOWE + Verb phrases | 1.025 | 0.458 |
| | SIM+Verb phrases | 1.016 | 0.488 |
| GloVe (Metzler, 2019) | SOWE + Verb phrases | 1.036 | 0.425 |
| | SIM+Verb phrases | 1.002 | 0.509 |
| FastText (Metzler, 2019) | SOWE + Verb phrases | 1.023 | 0.465 |
| | SIM+Verb phrases | 0.956 | 0.537 |
| ELMo | SIM | **0.997** | **0.482** |
| BERT | SIM | 1.058 | 0.322 |
| GPT-2 | SIM | 1.078 | 0.28 |

Table 2: Results on Mohler dataset with various models/approaches.[18]

## 7. Ethical Considerations

LLMs graders have emerged as efficient solutions for grading student performance. However, integrating them raises ethical considerations that must be carefully addressed.

The bias in the evaluation process is a significant ethical issue with automated evaluation systems. That's why Implementing strategies to recognize and correct it in the data or system will help to lessen this. This way, educational institutions can maintain the fairness and integrity of the evaluation process while ensuring equal opportunity for all students. In order to guarantee consistency and accuracy in the evaluation, it is also essential to precisely identify what the grading system is evaluating using well-defined criteria. This way, we will ensure a more meaningful and fair review. Moreover, clarity in the assessment standards fosters mutual respect and comprehension between instructors and students, guaranteeing the validity of the computerized assessment procedure.

Depending on the discipline, determining "correct" responses might be a challenging task. Compared to interpretive disciplines that involve nuanced perspectives, identifying measurements for correctness may be easier in STEM fields where issues have objective answers. When the model assesses fields like sociology, which don't have a single correct solution, ethical considerations may arise[17]. As we previously noted, grading systems with preset grading rubrics can be used to accurately grade objective STEM disciplines. However, in order to accurately assess different legitimate points of view, more qualitative matters call for human judgment. Educational institutions could use a variety of strategies, such as multiple graders evaluating sociology, partial credit allocation, adding discussion-based assignments to automated scoring, creating detailed rubrics based on interpretation, and continuously improving criteria based on feedbacks from students and educators.

Consideration of stakeholder roles, which are educators, administrators, students and developers, is also crucial for ethical automation. Educators must comprehend the model's limitations to skillfully monitor diverse automated outputs. However, there remains a risk that despite best efforts: an automated grading system could potentially give a mistaken mark. This problem raises important questions : Who is responsible for the model's mistake? And how to remedy such situation ethically?[2] While developers would address issues at a technical level, the educational institution holds the responsibility for guaranteeing the certification of learning achievements. They must provide transparent policies for how students can request re-evaluation. It is also crucial that automated grades be monitored by instructors who can catch outliers.

By prioritizing ethical considerations and implementing

these measures, we can leverage the benefits of automated evaluation while fostering a meaningful educational experience for all students.

## 8. Conclusion

We conducted an assessment of the embeddings produced by three distinct transfer learning models using the Mohler dataset, which is specific to the domain of Computer Science, for the ASAG task. These transfer learning models —ELMo, BERT, and GPT-2— were succinctly introduced along with their respective architectures.

The creation of sentence embeddings involved all three selected transfer learning models, encompassing both desired and student answers in the dataset. The encoding of answers is connected to the words within them, regardless of their sequence. Subsequently, the cosine similarity feature was computed for each student answer and desired answer. This feature underwent training using three distinct regression methods: linear, istonic and ridge.

ELMo emerged as the top-performing transfer learning model for the task, achieving a remarkable RMSE score of 0.980 and a Pearson correlation of 0.482.

Moreover, the integration of Language Model (LLMs) graders offers efficiency in student performance grading, yet demands careful consideration of ethical issues such as bias and clarity in evaluation criteria. Stakeholder roles, including educators, administrators, students, and developers, are pivotal in ensuring ethical automation. By addressing these concerns and prioritizing ethical considerations, we can harness the advantages of automated evaluation while maintaining a meaningful and fair educational experience for all students.

## 9. Future Work

Our assessment of transfer learning models—ELMo, BERT, and GPT-2—on the ASAG task using the Mohler dataset reveals avenues for future exploration. One key area is to leverage the success of ELMo in feedback generation for the ASAG task. Fine-tuning these models specifically for feedback purposes could yield valuable advancements in automated assessment and instructional support.

Extending evaluations to diverse datasets and educational domains would provide insights into model generalizability. Exploring alternative machine learning techniques beyond linear and ridge regression, such as neural network-based approaches, could enhance feedback generation. Additionally, investigating methods to improve the interpretability of generated feedback and integrating real-time feedback tools into educational platforms are worthwhile directions for future research.


## 10. Acknowledgments

We would like to express our sincere gratitude to M. Llored, M. Moulla and M. EL Rhabi, for their invaluable guidance and support throughout the process of writing this article. Their expertise and thoughtful feedback greatly enriched the content.


## References


[1] T.R. Akila Devi, K. Javubar Sathick, and A. et al. Abdul Azeez Khan. "Novel Framework for Improving the Correctness of Reference Answers to Enhance Results of ASAG Systems". In: *SN Computer Science* 4 (2023), p. 415. DOI: 10.1007/s42979-023-01682-8. URL: https://doi.org/10.1007/s42979-023-01682-8.

[2] Christoph Bartneck et al. *An Introduction to Ethics in Robotics and AI*. Jan. 2021. ISBN: 978-3-030-51109-8. DOI: 10.1007/978-3-030-51110-4.

[3] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *CoRR* abs/1810.04805 (2018). arXiv: 1810.04805. URL: http://arxiv.org/abs/1810.04805.

[4] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: (2019"). arXiv: 1810.04805 [cs.CL].

[5] Hadi Abdi Ghavidel. "Automatic Short Answer Grading Using Transformers". MA thesis. Polytechnique Montréal, Feb. 2021. URL: https://publications.polymtl.ca/5608/.

[6] Ehab Hamdy. *Neural Models for Offensive Language Detection*. 2021. arXiv: 2106.14609 [cs.CL].

[7] Huongdo. *GitHub - huongdo108/sentiment-analysis-LSTM-BERT-XLNet*. https://github.com/huongdo108/sentiment-analysis-LSTM-BERT-XLNet. Accessed: [insert date here].

[8] Joddiy. *NLP Summary*. https://joddiy.github.io/2020/06/10/NLP-Summary/. Accessed: [Insert Access Date Here]. 2020.

[9] Thomas D. Metzler. "Computer-assisted Grading of Short Answers using Word Embeddings and Keyphrase Extraction". In: (2019).

[10] Michael Mohler, Razvan Bunescu, and Rada Mihalcea. "Learning to Grade Short Answer Questions using Semantic Similarity Measures and Dependency Graph Alignments". In: (June 2011). Ed. by Dekang Lin, Yuji Matsumoto, and Rada Mihalcea, pp. 752–762. URL: https://aclanthology.org/P11-1076.

[11] Michael Mohler and Rada Mihalcea. "Text-to-text Semantic Similarity for Automatic Short Answer Grading". In: *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*. 2009, pp. 567–575.

[12] *Papers with Code - ELMo Explained*. https://paperswithcode.com/method/elmo. Accessed: [insert date here].

[13] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. "GloVe: Global Vectors for Word Representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2014, pp. 1532–1543.

[14] Daniel Pérez et al. "About the Effects of Combining Latent Semantic Analysis with Natural Language Processing Techniques for Free-text Assessment". In: *Revista Signos* 38.59 (2005), pp. 325–343.

[15] Matthew E. Peters et al. *Deep contextualized word representations*. 2018. arXiv: 1802.05365 [cs.CL].

[16] Alec Radford et al. *Language Models are Unsupervised Multitask Learners*. 2019. URL: https://api.semanticscholar.org/CorpusID:160025533.

[17] D. Ramesh and S.K. Sanampudi. "An Automated Essay Scoring Systems: A Systematic Literature Review". In: *Artificial Intelligence Review* 55 (2022), pp. 2495–2527. DOI: 10.1007/s10462-021-10068-2. URL: https://doi.org/10.1007/s10462-021-10068-2.

[18] Paul G. Plöger Sasi Kiran Gaddipati Deebul Nair. "Comparative Evaluation of Pretrained Transfer Learning Models on Automatic Short Answer Grading". In: (2020). arXiv: 2009.01303 [cs.CL]. URL: https://arxiv.org/pdf/2009.01303.pd.

[19] Md Arafat Sultan, Cristobal Salazar, and Tamara Sumner. *Fast and Easy Short Answer Grading with High Accuracy*. Ed. by Kevin Knight, Ani Nenkova, and Owen Rambow. San Diego, California: Association for Computational Linguistics, June 2016", pp. 1070–1075. DOI: 10.18653/v1/N16-1123. URL: https://aclanthology.org/N16-1123.

[20] Mithun Thakkar. "Finetuning Transformer Models to Build ASAG System". In: (Sept. 2021).