Unsupervised Automatic Short Answer Grading and Essay Scoring: A Weakly Supervised Explainable Approach

Felipe Urrutia[1] **Cristian Buc** **Roberto Araya**[2] **Valentin Barriere**[1]
[1]Universidad de Chile, Department of Computer Science (DCC), Santiago, Chile
[2]Universidad de Chile, Centro de Investigación Avanzada en Educación (CIAE), Santiago, Chile
furrutia@dim.uchile.cl, cristian.buc@cenia.cl,
roberto.araya.schulz@gmail.com, vbarriere@dcc.uchile.cl

Abstract

Automatic Short Answer Grading (ASAG) refers to automated scoring of open-ended textual responses to specific questions, both in natural language form. In this paper, we propose a method to tackle this task in a setting where annotated data is unavailable. Crucially, our method is competitive with the state-of-the-art while being lighter and interpretable. We crafted a unique dataset containing a highly diverse set of questions and a small amount of answers to these questions; making it more challenging compared to previous tasks. Our method uses weak labels generated from other methods proven to be effective in this task, which are then used to train a white-box (linear) regression based on a few interpretable features. The latter are extracted expert features and learned representations that are interpretable *per se* and aligned with manual labeling. We show the potential of our method by evaluating it on a small annotated portion of the dataset, and demonstrate that its ability compares with that of strong baselines and state-of-the-art methods, comprising an LLM that in contrast to our method comes with a high computational price and an opaque reasoning process. We further validate our model on a public Automatic Essay Scoring dataset in English, and obtained competitive results compared to other unsupervised baselines, outperforming the LLM. To gain further insights of our method, we conducted an interpretability analysis revealing sparse weights in our linear regression model, and alignment between our features and human ratings.[1]

1 Introduction

Applications of Large Language Models (LLMs) are emerging in the field of education and have taken complementary roles to those of teachers (Jeon and Lee, 2023). For instance, LLMs have been used, with mixed results, to train teachers to learn new strategies (Wang and Demszky, 2023). One aspect of education that can greatly benefit of automation is that of grading or scoring (Lan et al., 2024). Such automation could greatly improve the flexibility of teaching and target on the fly specific educational shortcomings of students.

In this work, we focus on two of these automations: *(i)* **Automatic Short Answer Grading (ASAG)** and *(ii)* **Automatic Essay Scoring (AES)**; both instances of automated scoring for open-ended questions. More specifically, ASAG focuses on grading short, open-ended responses. These responses are typically a few sentences to a paragraph long and are often fact-based, requiring concise answers. In contrast, AES evaluates longer, more complex pieces of writing, which typically contain an introduction, body, and conclusion, and involve argumentation, analysis, and critical thinking. AES is one of the earliest research problems in natural language processing (Page, 1966, 1967).

One crucial aspect of automated grading on open-ended questions is the ability to interpret the grade. The machine learning community has prioritized increasing explainability in models, leading to the emergence of Explainable AI. This area focuses on building tools to understand the decisions made by learning models (Gunning et al., 2019; Arrieta et al., 2020; Fel et al., 2022), or even advocates for the sole use of white-box models (Rudin, 2019). However, white-box models typically display poorer performances compared with black-box ones (Loyola-Gonzalez, 2019). Thus, in line with the explainable trend, recent methods have focused on developing novel tools to increase the performance of white-box models, sometimes up

[1]Code available at furrutiav/unasages-bea2025

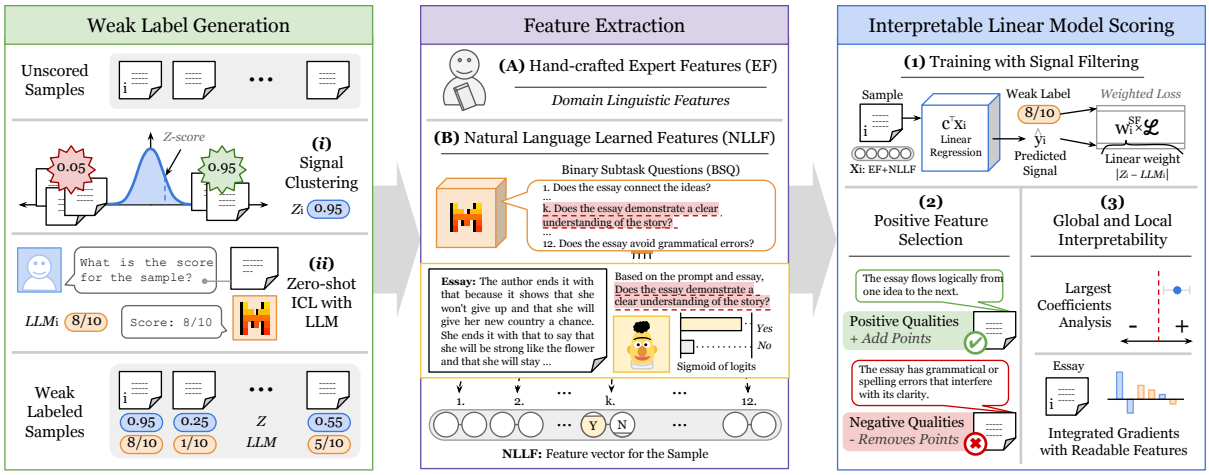


Figure 1: Full process. **Phase 1**, Generation of Weak Labels using Unsupervised methods: Signal Clustering (Chen et al., 2010a) or through an LLM (Jiang et al., 2023b). **Phase 2**, domain Expert Features (EFs) extraction and Natural Language Learned Features (NLLFs) obtained from answers to Binary Subtask Questions (BSQs) (Urrutia et al., 2023). **Phase 3**, feature selection, interpretable model training and analysis.

to that of black-box models (Urrutia et al., 2023).

Finally, most studies rely on supervised learning with annotated datasets (Takano and Ichikawa, 2022; Bonthu et al., 2023; Zhang et al., 2022), where a few items are associated to many annotations. A situation that is barely encountered in real-life scenarios. Moreover, only a few works in this area focus on non-English language (Latif et al., 2024). The majority of them are restrained to English, and none of them in (Latin-American) Spanish.

Motivation and Contributions. In this work, we tackle the issues raised above in a single framework (see Figure 1). First, we propose a method that allows us to reach high performance in ASAG and AES tasks in an unsupervised way. Second, we show the potential of our model to create interpretable white-box predictions based on sparse features, in a setting where strong generalization abilities are required because of highly diverse questions with a few answers.

Therefore our contributions are as follows: *(i)* we present a novel Non-English language dataset that is particularly challenging for ASAG systems, as it involves many questions with few answers, *(ii)* we propose a novel framework that unifies unsupervised and supervised methods into a single ASAG/AES system. In particular, we use weak labels from opaque unsupervised methods for supervised learning in white-box models, *(iii)* we propose a way to maximize the impact of the best-labeled training examples by weighting the loss function regarding the degree of consensus between each weak label, *(iv)* we compare our method with strong ASAG and AES baselines on two distinct datasets of different languages, and show that our method significantly outperforms previous white-box models, and falls barely short to LLM-based ASAG or to SOTA AES, *(v)* we run a thorough analysis on the AES dataset to demonstrate the interpretability of our method by: looking at our model's sparse weights, comparing it with SOTA using their integrated gradients but also showing our features are aligned with humans scores.

2 Related Work

In the context of ASAG, several methods have been proposed. Recent work has focused on generating understandable scoring by decomposing items (i.e., questions and responses to math problems) into rubrics whose validity can be inferred with language models (Hellman et al., 2023). Similar work have focused on directly fine-tuning pre-trained language models for ASAG (Takano and Ichikawa, 2022; Bonthu et al., 2023; Zhang et al., 2022), or training language models only based on student responses (Steimel and Riordan, 2020). Some works developed a hybrid ASAG system that evaluates answers to mathematical questions based on deterministic methods and the quality of explanations using text-based scoring methods (Cahill et al., 2020). Note that many semi-supervised (Brooks et al., 2014; Weegar and Idestam-almquist, 2024; Basu et al., 2013) or similarity-based methods (Bexte

et al., 2023) allow to use less labels, but they still need some of them.

In the context of AES, Taghipour and Ng (2016) were pioneers in training neural networks for AES, using a CNN-LSTM on the Automated Student Assessment Prize (ASAP; Hamner et al. 2012) dataset. Even though supervised models remain the most efficient (Yang et al., 2020), unsupervised methods like the one we are proposing show promising results. For instance, AESPrompt (Tao et al., 2022) obtains competitive results in one-shot essay scoring using continuous prompt learning. Wang et al. (2023) created a fully unsupervised approach using heuristic signals learning as a proxy task, as ultimate goal to train a BERT-based essay scorer, and obtained state-of-the-art performances on ASAP. Recent works have focused on the ability of LLMs to automatically score the proficiency of written essays on ASAP (Mansour et al., 2024; Lee et al., 2024). Stahl et al. (2024) even proposed prompting strategies for joint essay scoring and feedback generation to gain more interpretability.

Regarding general explanability, techniques that could be used for ASAG and AES such as Chain-of-Thought (CoT) (Wei et al., 2022) can provide a superficial level of explanation but are prone to structural biases in the text that put in question their fidelity (Turpin et al., 2023; Paul et al., 2024). Moreover, these techniques are fragile as pre-trained language models show lack of robustness on adversarial or unusual writing (Lottridge et al., 2023). Importantly, these writing types are often present in the answers of young children like in the ASAG dataset of this study.

3 Methods

The task of automatically assigning scores to short answers/essays involves finding a model M that assigns a score $\hat{y}_i$ between 1 and $S_{\max}$ to each pair of question/answer or instruction/essay. **First**, we use unsupervised methods to create weak labels. **Second**, we represent every document using interpretable features. **Third**, we select features and train a non-negative linear regression model on the weak labels, using a special loss to maximize the weak labels quality. We show the model is both white-box, sparse and interpretable.

Weak-supervision We propose to train an unsupervised model M by leveraging high-level heuristic signals, or weak labels. Our method (see Figure 1, **Phase 1**) involves utilizing two distinct signals: *(i)* scores derived from the unsupervised Signal Clustering method (SC; Chen et al. 2010b, see below) and *(ii)* scores obtained from an LLM using zero-shot in-context learning. For a given question-answer/instruction-essay pair (q_i, a_i), we denote as Z_i the signal of the answer with SC or LLM_i the LLM-based signal. To weakly-supervise the training of M, we use $y_i = Z_i$ or $y_i = \text{LLM}_i$ in order to minimize the loss function $\mathcal{L}(\hat{y}_i, y_i)$.

Signal Clustering (or Z-score) Based on Chen et al. (2010a), this method is simple yet allows for surprisingly good results in unsupervised automatic essay scoring. Basically, it initialize each essay score with a simple value, and then iteratively propagates the scores to other samples in the same cluster. For their essay scoring task, the authors of the original paper used the number of unique terms in the answer as the initial score. It uses the following inductive formula:

$$
\begin{aligned}
Z_{i0}: &\quad \text{Initial score for the } i\text{-th answer,} \\
S_{it} &= \sum_{j \neq i} \text{Sim}_{ij} \cdot Z_{i(t-1)}, \\
Z_{it} &= \frac{S_{it} - \frac{1}{N-1} \sum_{k \neq i} S_{kt}}{\sigma_t},
\end{aligned}
$$

where S_{it} is the score for the i-th answer at step t, Sim_{ij} is the similarity between the i-th answer and the j-th answer, and Z_{it} is the Z-score of the i-th with σ_t the standard deviation of $S_{\cdot t}$ at step t. We call Z_i the Z-score of the i-th answer at final step. We update Z_i until convergence.

Interpretable Features Following the work of Urrutia et al. (2023), we incorporated a set of expert-derived features (EF) coming from expert domain knowledge, and also high-level explainable features such as Natural Language Learned Features (NLLF; Urrutia et al. 2023). NLLFs encode answers to simpler-than-the-task binary questions, called **Binary Subtask Questions (BSQs)**, into a human-readable feature vector. It allows the model to represent each sample as a vector of probabilities on other interpretable simpler sub-tasks, like "*Is the answer written clearly and concisely?*". More details are available in (Urrutia et al., 2023) and in Appendix B. We also use the concatenation of both type of features (EF+NLLF). For EF, we use in ASAG/AES a list of 36/14 hand-crafted features, to describe the answers to math questions/essays (Table 6/7 in Appendix). Figure 1 shows the feature

Question	Answers	Score
Don Antonio bought 3 boxes of cereal at $673 each. The seller charged him $2100. Is what they charged him correct? Explain in your own words.	If Don Antonio bought 3 boxes, it's fine. No, because he should be charged less. It's not wrong, I got 2019.	{2, 3} {4, 3} {6, 7}

Figure 2: Examples of a Question, Answers and Scores from our ASAG dataset. Translated from Spanish.

extraction in **Phase 2**, with an example of BSQ and the NLLF vector for an essay.

Interpretable Model: Linear Regression We trained a linear regression on two types of weak labels (see Figure 1, **Phase 3**).

Signal Filtering We propose a method to maximize the impact of well-labeled examples through the weighting of the loss function with respect to the degree of consensus among weak labels (see Figure 1, **Phase 3**). Basically, we compute linear weights utilizing the difference between the predicted scores generated by the LLM and the ones derived from the Signal Clustering method, both of which obtained in a unsupervised way. For a given question-answer pair (q_i, a_i) and weak-label $y_i \in \{Z_i, \text{LLM}_i\}$, the weighted loss is $w_i^{\text{SF}} \cdot \mathcal{L}(\hat{y}_i, y_i)$, where:

$$w_i^{\text{SF}} = 1 - \frac{|Z_i - \text{LLM}_i|}{S_{\max} - S_{\min}}.$$

Feature Selection In order to keep our model interpretable, we used two tricks (see Figure 1, **Phase 3**). First, we only chose BSQs formulation that were positively correlated with the score of the student[2] i.e., describing events that were seen as positive by the teacher. Second, we forced the linear regression model to learn only positive weights (Slawski and Hein, 2013) as they are applied on features that are positives w.r.t. the score. Section 5 shows that this setting allows for sparsity in the parameters space of the linear regression model.

4 Experiment and Results

4.1 Datasets and Evaluation Metrics

We ran experiments on two distinct tasks using two datasets in different languages. The first set of experiments (Section 4.1.1) tackles ASAG in Spanish while the second set of experiments (Section 4.1.2) tackles AES in English.

[2]using weak labels

Task	Genre	Avg. Length	Score Range	# Essays
1	PER	350	2-12	1783
2		350	1-6	1800
3		150	0-3	1726
4	SDE	150	0-3	1772
5		150	0-4	1805
6		150	0-4	1800
7	NAR	250	0-30	1569
8		650	0-60	723

Table 1: Properties of the different tasks in the AES dataset called ASAP. Genre: PER (persuasive), SDE (source-dependent), NAR (narrative).

4.1.1 Automatic Short Answer Grading in Spanish

The dataset comprises written answers from fourth-grade students to mathematics questions. The question-answers pairs were collected using the online e-learning platform ConectaIdeas, which is currently deployed and use by teachers and students in Chile. Its data was already used in past scientific studies (Urrutia Vargas and Araya, 2023; Urrutia and Araya, 2023). It encompasses a total of 63,612 answers to 1,248 unique questions collected across two academic years. The answers were obtained from a total of 3,463 fourth-grade students, with 231 for the 2017 period and 3,232 for the 2022 period. The answers have on average a total of 50 characters. Each question has on average a total of 52 answers per question for 2022 and 30 for 2017.

The data are annotated based on the scoring of answers for one academic year (2017). Answers from the unlabeled academic year are utilized to train automatic systems, while those from the labeled academic year serve as a test set for evaluating the performance of these systems. Annotation was conducted by two elementary mathematics teachers, assigning scores ranging from 1 to 7 (i.e., from insufficient to excellent). Only the scores from one teacher were utilized as the ground-truth, while the scores from the other teacher were utilized to analyze human performance, in this sense we can make a model that predicts the grading behavior of one teacher. We calculate the agreement between their scores and obtained a Correlation of

Method	Weak Signal	Signal Filt.	Text	EF	NLLF	EF + NLLF
Length	None	-	.2734	-	-	-
Jaccard Sim.	None	-	.2758	-	-	-
Cosine Sim.	None	-	.3759	-	-	-
ULRA	LF	-	.5112	-	-	-
	EF	-	.4264	-	-	-
	EF+LF	-	.4218	-	-	-
Z-score	None	-	.5104	-	-	-
LLM	None	-	**.5727**	-	-	-
LLM-CoT	None	-	.4778	-	-	-
Linear Regression	Z-score	✗	-	.4937	.3853	.5096
	LLM-based signal	✗	-	.4815	.3538	.4312
	Z-score	✓	-	.5018	.3899	**.5450**
	LLM-based signal	✓	-	.4974	.3712	.4791
BERT	Z-score	✗	.5220	-	-	-
	LLM-based signal	✗	.5085	-	-	-
	Z-score	✓	.5280	-	-	-
	LLM-based signal	✓	.5430	-	-	-
Human	None	-	**.7568**	-	-	-

Table 2: Results of the ASAG models using Pearson correlation: the cheap baselines using similarity, the ULRAs using different weak linguistics signals, the Z-score and LLM predictions, and our weakly supervised models. For the weakly supervised models, the linear model utilizes all combinations of two feature sets (EF and NLLF), while the BERT model is trained on text data.

.76. Figure 2 shows an example of the dataset.

4.1.2 Automatic Essay Scoring in English

We ran experiments using the Automated Student Assessment Prize[3] (ASAP) dataset (Hamner et al., 2012). This dataset has been widely used in several AES studies (Xie et al., 2022; Jiang et al., 2023b; Muangkammuen and Fukumoto, 2020; Mansour et al., 2024; Mathias and Bhattacharyya, 2018). For instance, it has been used by Wang et al. (2023) to assess the ULRA model for an usupervised AES task. It is composed of 12,978 essays divided into 8 different sets. Each of the sets corresponds to a specific essay task or prompt, which can be seen as domain. The tasks are of different genres: persuasive, source-dependent response, and narrative. The statistics of the dataset is shown in Table 1.

As a validation metrics, we report Quadratic Weighted Kappa (QWK) in order to compare the different models, generally utilized to measure the agreement between groundtruth scores and predicted scores on this dataset and in AES research.

4.2 Baselines

Dummy Baseline We use a regression model based on the answer length in terms of number of characters.

Similarity Measures We calculate the similarity between the question and the answer to assess its correctness based on the shared information between them.We use two methods: Jaccard Similarity on sparse embeddings (Bag-of-Words; (Harris, 1954)), and cosine similarity with dense vectors obtained from the `[CLS]` token of a multilingual Sentence Transformer (Reimers and Gurevych, 2019).

Signal Clustering (Z-score) Based on Chen et al., we use answer length as the initial scoring and assessed answer similarity based on the shared terms between two answers.

Mixtral We used a recent LLM to address the task in a zero-shot format (Jiang et al., 2023a), using a simple prompt containing the definition of the task. More details in Appendix C.

ULRA We implemented the unsupervised ULRA method of Wang et al. (2023) which showed state-of-the-art results on Automated Essay Scoring, which is close to ASAG. This model consists of using multiple quality signals obtained from heuristics as the pseudo-groundtruth, and then training a neural model by learning from the aggregation of these signals. The idea is that the final score should depend on an aggregation of these simple signals. For the ASAG task, we adapt the method translating the original Linguistic Features (LF) to Spanish, and by using our own Expert Features (EF) as pseudo-groundtruth. For the AES task, we utilized the LF from the original paper on the same task. Note that ULRA was also considered as a

[3] https://www.kaggle.com/c/asap-aes

Method	Weak Signal	Signal Filt.	Text	EF	NLLF	EF + NLLF
Length	None	-	.3893	-	-	-
Jaccard Sim.	None	-	-.1821	-	-	-
Cosine Sim.	None	-	.0237	-	-	-
ULRA (Wang et al., 2023)	LF	-	**.6423**	-	-	-
Z-score (Chen et al., 2010b)	None	-	.5809	-	-	-
LLM (Jiang et al., 2023a)	None	-	.5119	-	-	-
LLM-CoT (Wei et al., 2022)	None	-	.4152	-	-	-
MTS† (Lee et al., 2024)	None	-	.550	-	-	-
Linear Regression	Z-Score	✗	-	.5528	.6083	.6141
	LLM-based signal	✗	-	.5762	.5720	.6385
	Z-score	✓	-	.5603	.6123	.6255
	LLM-based signal	✓	-	.5797	.5814	**.6451**
BERT	Z-Score	✗	.5728	-	-	-
	LLM-based signal	✗	.4418	-	-	-
	Z-score	✓	.5764	-	-	-
	LLM-based signal	✓	.4781	-	-	-
AES-Prompt† (one-shot) (Tao et al., 2022)	None	-	.639	-	-	-
R^2-BERT† (supervised) (Yang et al., 2020)	None	-	.794	-	-	-
Human	None	-	**.7384**	-	-	-

Table 3: Results of the models on the AES task using the average of the QWK over the different essay tasks. We report the cheap baselines using similarity, ULRAs using different weak linguistics signals, the Z-score and LLM predictions, and our weakly supervised models. Human scores were re-calculated here. † From original papers.

weak label generation method, but did not generate favorable results.

Weakly supervised BERT We evaluated different BERT models (Devlin et al., 2019) with a regression head on top of the `[CLS]` vector to predict the weak signals. For the ASAG task, we used BETO, a Spanish BERT transformer (Cañete et al., 2023). For the AES task, we used an English BERT.[4]

Multi-trait Specialization We compare with the work of Lee et al. (2024), who proposed an unsupervised method using LLMs to predict the quality of essays in a zero-shot way. Their method learns to decompose the writing proficiency into distinct traits, as some are known to be useful for judging global essay quality (Ke and Ng, 2019) such as *Position* and *Thesis Clarity*, *Organization and Structure* or *Supporting Details and Evidence*.

4.3 Experimental Protocol

The `transformers` library (Wolf et al., 2019) was used to access the pre-trained model and to train our models. We used BETO as backbone for NLLF generation, and the 4-bit version of `Mixtral-8x7b`[5] as LLM. The linear regressions were trained using scikit-learn (Pedregosa et al., 2012). We standardized every features before the logistic regression. For the ASAG task, Pearson correlation measured the correlation between predicted scores from automatic models and ground-truth scores from one teacher. We evaluated our model on the 1,315 manually annotated examples. For the AES task, we randomly split the data into a training, a validation and a test sets following the proportion 60/20/20 like Wang et al. (2023).

4.4 Results

4.4.1 Results on ASAG

Table 2 shows the results of the different baselines and models. It is notable that naive baselines like a linear regression using the answer length can reach a correlation of .27, and are surpassed by similarity between answer and question using a sentence-bert. Best machine results (.57) are obtained with an LLM, surprisingly without using the CoT mechanism, but still far away from human performances (.75). All our weakly supervised approaches benefit from the Signal Filtering method. Adding NLLF to our method helps when using Z-value or the LLM output as weak label, allowing to reach a score close to the one of the LLM, but with an interpretable white-box algorithm (contrary to BERT). ULRA methods, using general and/or domain expert features, tend to display lower scores when compared with Signal Clustering and remaining methods in this task. Finally, the scores of the BERT model trained with the weak-labels are im-

[4] `bert-base-cased`, `bert-base-spanish-wwm-cased`
[5] `mistralai/Mixtral-8x7B-Instruct-v0.1`

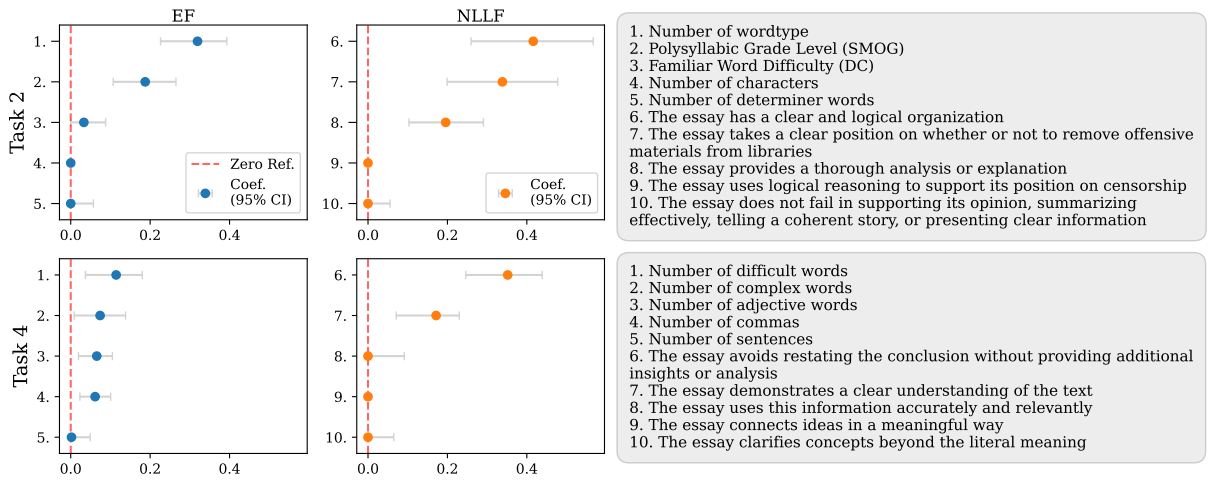


Figure 3: Highest coefficients of the Linear Regression with Signal Filtering using EF+NLLF on AES tasks 2 and 4. The box represents the 95% confidence interval. Biases are respectively of 3.37 and 1.35 for tasks 2 and 4.

proved when applying Signal Filtering, with the best one of .543 using an LLM-based signal as weak labels.

4.4.2 Results on AES

Table 3 shows the results of the different baselines and models on the AES task. Simple baselines achieve a moderate correlation (e.g., .4785 when using the answer length), while basic similarity measures, such as Jaccard and Cosine, perform poorly, with negative or near-zero correlations. Among the models, our method achieves the highest score (.645), outperforming other methods such as ULRA (.642), Z-score (.581), and LLM (.512), though all falling short of human-level performance (.738). Interestingly, the LLM with a CoT approach performs worse than the standard LLM, with a correlation of only .415, which is unexpected given the reported success of CoT in other contexts, specially for a task such as essay scoring in English. Notably, all of our weakly supervised models benefit significantly from the Signal Filtering method. Furthermore, adding the NLLF mechanism further enhances performance. Indeed, combining LLM-based labels, Signal Filtering, and NLLF reaches reaches the highest performance, outperforming prompt engineering baselines such as MTS or AES-Prompt. Finally, the BERT models trained with the weak-labels display lower scores (highest BERT score of .577 using Z-score and Signal Filtering). As a way to cross-check our results, existing works assessed the capacity of various LLMs on this tasks and dataset (Mansour et al., 2024; Lee et al., 2024). The performances we obtained (QWK of 0.51), are in line with the ones reported in Lee et al. (2024)[6], but higher than the ones reported in Mansour et al. (2024).

5 Analysis

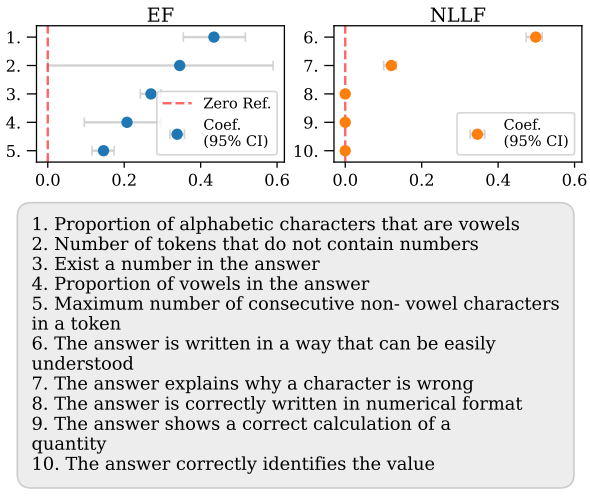


Figure 4: Highest coefficients of the Linear Regression with Signal Filtering using EF+NLLF features (Table 9). The box represents the 95% confidence interval. Bias is 4.65.

ASAG Coefficients Our best linear model uses a combination of only 6 coefficients: 4 hand-crafted features (EF) and 2 natural language learned features (NLLF). Figure 4 shows, from the most relevant features, that correct answers require a balanced use of vowels[7] (Features 1 and 5) or numbers

[6] 0.48 with a Mistral-7b-instruct

[7] Words with a balanced vowel-consonant structure, like the CVCVC pattern, are easier for children to process

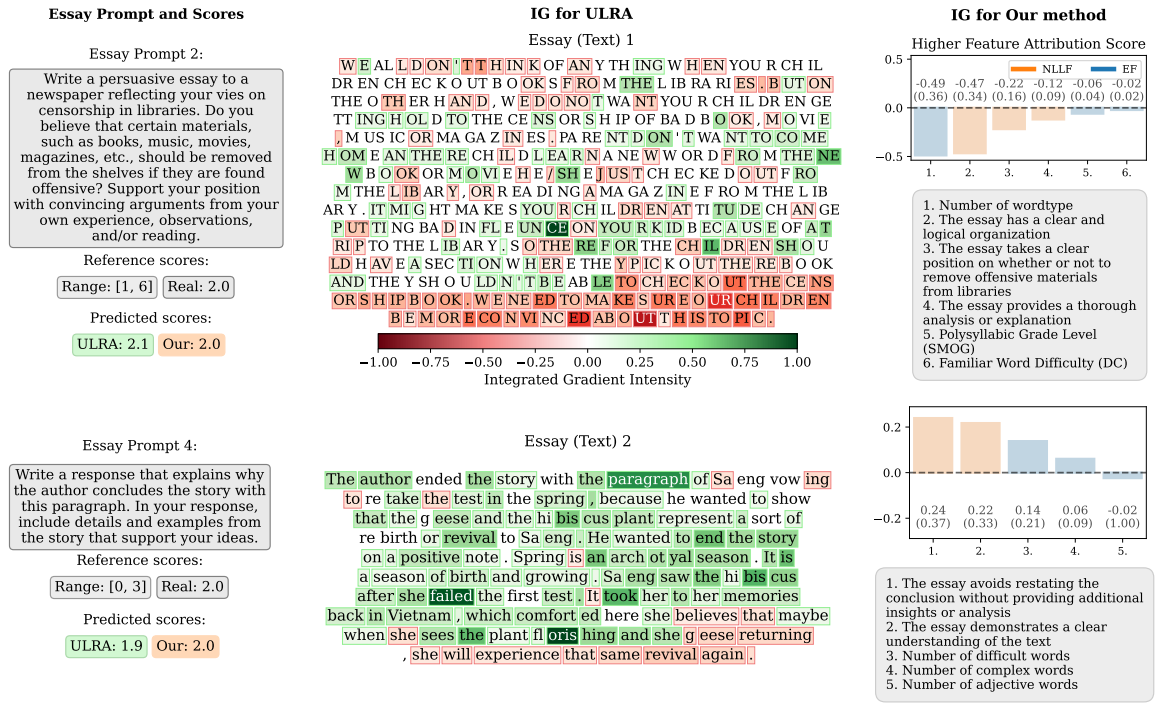


Figure 5: IG feature attribution examples from our method and ULRA on tasks 2 and 4 of the AES dataset.

(Feature 2, 3). In addition, NLLFs address common questions in which students are asked to explain if a character is making or not the right choice (Feat. 7) or just if the answer is clear (Feat. 6).

AES Coefficients Figure 3 show the coefficients of two of the eights linear regression models trained using NLLF+EF, respectively on tasks 2 and 4, persuasive and source-dependent genre, respectively. Most features have coefficients equal to zero, making the linear regression model very sparse, and leaving six usable features for each of the two models: 3 EF and 3 NLLF for the task 2, and 4 EF and 2 NLLF for the task 4. For the persuasive task, NLLFs are about argumentative techniques of the writer, whether or not it takes strongly position, and the structure of the essay.

AES IG Interpretability We claim that our system is white-box, but also interpretable. To back up our claim, we compare the two best performing models with a classical interpretation technique using Integrated Gradients (IG; Sundararajan et al. 2017) in order to attribute a score to each input feature. Figure 5 shows examples of feature attribution comparing our method and ULRA[8]. Whereas the attribution from the IG is complex to analyze in ULRA, our method offers two interesting advantages: (i) it is simple to interpret as it has only a few parameters which are all described in natural language, (ii) it identifies whether essays offer clear analyses or lack clear stances.

AES Human Interpretability We designed two experiments to manually validate our claim that NLLF values are coherent with humans judgments. First, we manually annotated 171 examples w.r.t the BSQ labels, in order to estimate the performances of the LLM and the NLLF Generator (NLLFG) in the subtasks. We find that both the LLM and the NLLFG obtain satisfying accuracies of .89 and .84, in concordance with the analysis of Urrutia et al. (2023). Second, for each BSQ, we selected pairs of examples based on deciles in the normalized distribution of the BSQ NLLF values. Each pair came from examples separated either by high (9 bins), medium (5 bins), or low (1 bin) distances in the distribution. We asked a human to annotate for each pair of examples, the one with highest NLLF value and the bin distance between the examples of the pair. This rendered a 6-class ordinal problem with 171 pairs. We obtained an accuracy of .44 (random is .16), an accuracy with a tolerance of 1 (Gaudette and Japkowicz, 2009) of .77 (random is .44) and a Krippendorff (2013)'s α

(Jiménez González and Garcia, 1995; Brame, 1974) and help recognize proper words like a measure of coherence (Urrutia Vargas and Araya, 2023).

[8]For the linear regression, the integrated gradient is simply the product between the feature and its weight.

of 0.63 (random is 0). More details in Appendix F.

6 Conclusion and Future Work

In unsupervised ASAG of young students to diverse open ended questions in Spanish, and unsupervised AES in English, SoTA LLM-based methods are still far away from human performances. Moreover, the models trained in answer scores made with LLMs can be approximated by much simpler and interpretable models. Weak supervision on LLM labels but also on target values that are way simpler including Signal Clustering is a potential avenue of research for white-box model using several types of interpretable features such as the combination of linguistic-based expert-domain ones and compositionality-based learned ones. Future work should focus on more intensive search on the prompt space, as well as involve supervised learning (and not only weakly supervised learning) and out-of-distribution question analysis. Regarding the interpretability, the integrated gradients could be back-propagated up to the tokens in order to visualize the impact of each of them on each NLLF.

Limitations

Our work has been put in use in Spanish for a very specific type of questions that are from math exams, and in English essay with a higher quality of the text content. It would be interesting to try it in a multilingual setting, using multilingual LLMs. Future works would also imply weakly supervised multi-task learning, and more advanced prompt engineering such as the one of Lee et al. (2024), that allows for decomposing an essay into multiple traits to better score it using an LLM. Finally, it would be interesting to use manually crafted BSQs using the annotation guidelines instead of generating them, in order to see if it will improve the quality of the final model.

Ethics Statement

This work is in compliance with the ACL Ethics Policy as it allows to create models that might be more interpretable in a sensitive context such as young student education.

Acknowledgements

The authors would like to thank the Basal Funding for Centers of Excellence from ANID/PIA for their support through the Centro de Investigación Avanzada en Educación (CIAE) with grant number FB0003.

References

Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115.

Sumit Basu, Chuck Jacobs, and Lucy Vanderwende. 2013. Powergrading: a Clustering Approach to Amplify Human Effort for Short Answer Grading. *Transactions of the Association for Computational Linguistics*, 1:391–402.

Marie Bexte, Andrea Horbach, and Torsten Zesch. 2023. Similarity-Based Content Scoring - A more Classroom-Suitable Alternative to Instance-Based Scoring? In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 1892–1903.

Sridevi Bonthu, S. Rama Sree, and M. H.M. Krishna Prasad. 2023. Improving the performance of automatic short answer grading using transfer learning and augmentation. *Engineering Applications of Artificial Intelligence*, 123(April):106292.

Michael K Brame. 1974. The cycle in phonology: stress in Palestinian, Maltese, and Spanish. *Linguistic Inquiry*, 5(1):39–60.

Michael Brooks, Sumit Basu, Charles Jacobs, and Lucy Vanderwende. 2014. Divide and correct: Using Clusters to Grade Short Answers at Scale. In *Learning@Scale*, pages 89–98.

Aoife Cahill, James H. Fife, Brian Riordan, Avijit Vajpayee, and Dmytro Galochkin. 2020. Context-based automated scoring of complex mathematical responses. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 186–192.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2023. Spanish pre-trained bert model and evaluation data. *arXiv preprint arXiv:2308.02976*.

Yen Yu Chen, Chien Liang Liu, Chia Hoang Lee, and Tao Hsing Chang. 2010a. An unsupervised automated essay-scoring system. *IEEE Intelligent Systems*, 25(5):61–67.

Yen-Yu Chen, Chien-Liang Liu, Chia-Hoang Lee, Tao-Hsing Chang, et al. 2010b. An unsupervised automated essay-scoring system. *IEEE Intelligent systems*, 25(5):61–67.

Jacob Devlin, Ming-wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL*.

Thomas Fel, Lucas Hervier, David Vigouroux, Antonin Poche, Justin Plakoo, Remi Cadene, Mathieu Chalvidal, Julien Colin, Thibaut Boissin, Louis Bethune, Agustin Picard, Claire Nicodeme, Laurent Gardes, Gregory Flandin, and Thomas Serre. 2022. Xplique: A Deep Learning Explainability Toolbox. In *Workshop on Explainable Artificial Intelligence for Computer Vision (XAI4CV)*, pages 5–8.

Lisa Gaudette and Nathalie Japkowicz. 2009. Evaluation methods for ordinal classification. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 5549 LNAI:207–210.

David Gunning, Mark Stefik, Jaesik Choi, Timothy Miller, Simone Stumpf, and Guang-Zhong Yang. 2019. XAI—Explainable artificial intelligence. *Science robotics*, 4(37):eaay7120.

Ben Hamner, Jaison Morgan, Lynnvandev, Mark Shermis, and Tom Vander Ark. 2012. The Hewlett Foundation: Automated Essay Scoring. Kaggle.

Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Scott Hellman, Alejandro Andrade, and Kyle Habermehl. 2023. Scalable and explainable automated scoring for open-ended constructed response math word problems. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 137–147, Toronto, Canada. Association for Computational Linguistics.

Jaeho Jeon and Seongyong Lee. 2023. Large language models in education: A focus on the complementary relationship between human teachers and ChatGPT. *Education and Information Technologies*, 28(12):15873–15892.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023a. Mistral 7b.

Zhiwei Jiang, Tianyi Gao, Yafeng Yin, Meng Liu, Hua Yu, Zifeng Cheng, and Qing Gu. 2023b. Improving Domain Generalization for Prompt-Aware Essay Scoring via Disentangled Representation Learning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 12456–12470.

Juan E Jiménez González and Carmen R Haro Garcia. 1995. Effects of word linguistic properties on phonological awareness in Spanish children. *Journal of Educational Psychology*, 87(2):193.

Zixuan Ke and Vincent Ng. 2019. Automated essay scoring: A survey of the state of the art. In *IJCAI International Joint Conference on Artificial Intelligence*, volume 2019-Augus, pages 6300–6308.

Klaus Krippendorff. 2013. Content Analysis: An Introduction to Its Methodology. In *Content Analysis: An Introduction to Its Methodology*.

Yunshi Lan, Xinyuan Li, Hanyue Du, Xuesong Lu, Ming Gao, Weining Qian, and Aoying Zhou. 2024. Survey of natural language processing for education: Taxonomy, systematic review, and future trends. *arXiv preprint arXiv:2401.07518*.

Ehsan Latif, Gyeong-Geon Lee, Knut Neuman, Tamara Kastorff, and Xiaoming Zhai. 2024. G-SciEdBERT: A Contextualized LLM for Science Assessment Tasks in German. pages 1–9.

Sanwoo Lee, Yida Cai, Desong Meng, Ziyang Wang, and Yunfang Wu. 2024. Unleashing Large Language Models' Proficiency in Zero-shot Essay Scoring. In *Findings of ACL: EMNLP 2024*.

Susan Lottridge, Chris Ormerod, and Amir Jafari. 2023. Psychometric considerations when using deep learning for automated scoring. *Advancing Natural Language Processing in Educational Assessment*, page 15.

Octavio Loyola-Gonzalez. 2019. Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view. *IEEE access*, 7:154096–154113.

Watheq Mansour, Salam Albatarni, Sohaila Eltanbouly, and Tamer Elsayed. 2024. Can Large Language Models Automatically Score Proficiency of Written Essays? In *2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC-COLING 2024 - Main Conference Proceedings*, pages 2777–2786.

Sandeep Mathias and Pushpak Bhattacharyya. 2018. ASAP++: Enriching the ASAP automated essay grading dataset with essay attribute scores. In *LREC 2018 - 11th International Conference on Language Resources and Evaluation*, pages 1169–1173.

Panitan Muangkammuen and Fumiyo Fukumoto. 2020. Multi-task Learning for Automated Essay Scoring with Sentiment Analysis. *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: Student Research Workshop*, (2015):116–123.

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

Ellis B Page. 1967. Statistical and linguistic strategies in the computer grading of essays. In *COLING 1967 Volume 1: Conference internationale sur le traitement automatique des langues*.

Debjit Paul, Robert West, Antoine Bosselut, and Boi Faltings. 2024. Making Reasoning Matter: Measuring and Improving Faithfulness of Chain-of-Thought Reasoning. In *ACL*, pages 15012–15032.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2012. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.

Martin Slawski and Matthias Hein. 2013. Non-negative least squares for high-dimensional linear models: Consistency and sparse recovery without regularization.

Maja Stahl, Leon Biermann, Andreas Nehring, and Henning Wachsmuth. 2024. Exploring LLM Prompting Strategies for Joint Essay Scoring and Feedback Generation. In *19th Workshop on Innovative Use of NLP for Building Educational Applications, Proceedings*.

Kenneth Steimel and Brian Riordan. 2020. Towards Instance-Based Content Scoring with Pre-Trained Transformer Models. *AAAI 2020 - 34th AAAI Conference on Artificial Intelligence*, (Shermis):2015–2017.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *34th International Conference on Machine Learning, ICML 2017*, volume 7, pages 5109–5118.

Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. In *EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, pages 1882–1891.

Shunya Takano and Osamu Ichikawa. 2022. Automatic scoring of short answers using justification cues estimated by bert. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 8–13.

Qiuyu Tao, Jiang Zhong, and Rongzhen Li. 2022. AESPrompt: Self-supervised Constraints for Automated Essay Scoring with Prompt Tuning. In *Proceedings of the International Conference on Software Engineering and Knowledge Engineering, SEKE*, pages 335–340.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. NeurIPS, pages 1–14.

Felipe Urrutia and Roberto Araya. 2023. Who ' s the Best Detective ? LLMs vs . MLs in Detecting Incoherent Fourth Grade Math Answers. *arXiv*.

Felipe Urrutia, Cristian Calderon, and Valentin Barriere. 2023. Deep natural language feature learning for interpretable prediction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3736–3763, Singapore. Association for Computational Linguistics.

Felipe Ignacio Urrutia Vargas and Roberto Araya. 2023. Automatic detection of incoherent written responses to open-ended mathematics questions of fourth graders. *MDPI Systems*, pages 0–35.

Cong Wang, Zhiwei Jiang, Yafeng Yin, Zifeng Cheng, Shiping Ge, and Qing Gu. 2023. Aggregating Multiple Heuristic Signals as Supervision for Unsupervised Automated Essay Scoring. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 1:13999–14013.

Rose Wang and Dorottya Demszky. 2023. Is ChatGPT a Good Teacher Coach? Measuring Zero-Shot Performance For Scoring and Providing Actionable Insights on Classroom Instruction. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, (Bea):626–667.

Rebecka Weegar and Peter Idestam-almquist. 2024. Reducing Workload in Short Answer Grading Using Machine Learning. *International Journal of Artificial Intelligence in Education*, 34(2):247–273.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing.

Jiayi Xie, Kaiwei Cai, Li Kong, Junsheng Zhou, and Weiguang Qu. 2022. Automated Essay Scoring via Pairwise Contrastive Regression. In *Proceedings - International Conference on Computational Linguistics, COLING*, volume 29, pages 2724–2733.

Ruosong Yang, Jiannong Cao, Zhiyuan Wen, Youzheng Wu, and Xiaodong He. 2020. Enhancing automated essay scoring performance via fine-tuning pre-trained language models with combination of regression and ranking. In *Findings of the Association for Computational Linguistics Findings of ACL: EMNLP 2020*, pages 1560–1569.

Mengxue Zhang, Sami Baral, Neil Heffernan, and Andrew Lan. 2022. Automatic Short Math Answer Grading via In-context Meta-learning. *Proceedings of the 15th International Conference on Educational Data Mining, EDM 2022*.

A ASAG Dataset Statistics

We present a summary of the dataset in Table 4, including the total number of students, different questions and student answers. We added the average number of answers per question for each year.

Year	#Students	#Questions	#Answers	Avg. #Ans. per Question
2022	3,232	1,204	62,297	≈ 52
2017	231	44	1,315	≈ 30
Total	3,463	-	63,612	-

Table 4: Summary of students, total questions, total answers, and average answers per question across years.

B Features

B.1 EF

We manually designed linguistic features, detailed in Table 6, aimed at capturing structural, morphological, and statistical properties of student responses for ASAG task in Chilean Spanish. Given the unique characteristics of children's writing in a mathematical context, we categorize EF into six groups: *morphological features*, which analyze the presence of numbers, digit counts, and the ratio of numerical to non-numerical tokens, essential for evaluating arithmetic-based responses; *syntactic features*, focusing on tokenization, negation length, and the distribution of non-numeric tokens, which help assess the sentence structure typical of early learners; *lexical features*, which measure character frequencies to detect common patterns in children's spelling and word usage in Chilean Spanish; *structural features*, capturing answer length, repeated character sequences, and vowel/consonant distributions, which are indicative of fluency and coherence; *punctuation features*, which count and analyze punctuation marks, distinguishing between mathematical symbols (e.g., decimal points, equation signs) and non-mathematical punctuation that might indicate explanatory attempts; and *phonological features*, assessing vowel proportions relative to alphabetic characters to identify phonetic simplifications or spelling mistakes common in young learners.

For example, the phonological feature measuring the proportion of alphabetic characters that are vowels (Feature 1) distinguishes between responses like A1 (0.33) and A2 (0.52) to the same question, with A2 being more phonetically fluent (see Table 5). Similarly, syntactic complexity can be estimated through the number of tokens without digits (Feature 2), where a detailed explanation (12 tokens, A2) correlates with a higher score than a brief response (2 tokens, A1). Morphological traits such as the binary presence of a number (Feature 3) allow us to capture relevant numerical grounding in an answer; for instance, A2 includes a number and scores higher. Phonological depth is further captured by vowel density (Feature 4), where answers with higher vowel proportion (0.31) exhibit better coherence than sparse ones (0.25). Finally, structural complexity, such as the maximum number of consecutive non-vowel characters in a token (Feature 5), helps detect unnatural or noisy tokens, e.g. A1 has a high value (5) due to "Hkflg", suggesting incoherence, compared to A2's more natural phrasing (value of 2).

B.2 NLLF

Following the method outlined in Urrutia et al., we utilize a selected roughly 12% subset of the train-set to generate the NLLF. We ask to a Mixtral to generate a diverse pool of Binary Subtask Questions (BSQ) for our ASAG/AES task. A member of our research team manually removes irrelevant BSQ. We chose 12 binary questions through automatic selection via Agglomerative Clustering, taking the centroid. We automatically answer the selected binary questions on the portion of the train-set with the same LLM to teach a Spanish/English BERT model in answering to all the selected binary questions. We generated a total of 24 features from the sigmoid of the logits of the trained BERT to provide *Yes* or *No* answers to the 12 binary questions (Table 6), i.e. two features per binary question.

C LLM

We used a simple prompt containing the definition of the task. For the AES task, initially, we use an unspecified prompt to score answers, yet observed a tendency for the model to assign notably low scores to answers containing kid misspelling errors. Subsequently, we refined our prompt specifying "not penalize for spelling mistakes and focus on the intended meaning conveyed by the student's answer". This adjustment yielded enhancements in the performance of the LLM.

D ULRA as Weak Signal

In the AES dataset, the LLM performances are outperformed by the ones obtained using the ULRA method, which is unsupervised but also black-box.

Feature	Question (Q) and Answer (A1-A2)	Feature Value	Score
Proportion of alphabetic characters that are vowels	**Q:** Si Jose multiplica 150 veces 1 ¿Cuál sería su resultado? Explica	-	-
	A1: 150x1 es 51 *(Low vowel ratio)*	0.33	3.0
	A2: sería 150 porque 150 veces 1 sería 150 *(Higher vowel ratio)*	0.52	7.0
Number of tokens without numbers	**Q:** José compró 4 cajas de leche a $245 cada una. El vendedor le cobró en total $950. ¿Está correcto lo que le cobró el vendedor? Explica.	-	-
	A1: está bien *(short, lacks analysis)*	2	2.0
	A2: está mal la respuesta es 980 se multiplica 245 x4 y el resultado es 980 *(detailed reasoning)*	12	7.0
Exist a number in the answer	**Q:** Paulina tiene 16 lápices para repartir entre 4 amigas. Su mamá le dice a Paulina que le va a dar 5 lápices a cada amiga. ¿Es correcto lo que le dice su mamá?	-	-
	A1: no es mal porque no *(no number)*	0	3.0
	A2: la mamá está mal porque son 4 lápices para cada amiga *(includes number)*	1	7.0
Proportion of vowels in the answer	**Q:** Una manzana pesa 0,35 g, otra 0,251 g y la última 0,51 g. ¿Cuánto pesan entre las tres?	-	-
	A1: 150x1 es 51 *(low vowel ratio)*	0.25	4.0
	A2: sumo todas las manzanas es 1,111. y sumé 0,35 más 0,251 más 0,51 y me dio ese resultado *(more fluent)*	0.31	7.0
Max. consecutive non-vowel characters	**Q:** ¿Cuál es el resultado de 501x2? Comenta cómo resolviste el ejercicio y explica qué es la multiplicación.	-	-
	A1: Hkflg *(noisy token)*	5	1.0
	A2: es 1 002, lo resolví con sumas y la multiplicación es una suma repetida *(coherent)*	2	7.0

Table 5: Examples of five expert features with their feature values for question/answer pairs in the ASAG task in Chilean Spanish (examples presented in their original language, Spanish).

For these reasons, we propose an additional experiments where we train a logistic regression model on our interpretable vector of expert and natural language learned representations, using the scores from ULRA as a weak label. The results are shown in Table 8. We can see that the use of a more accurate weak signal does not allow to improve the global performances.

E Prompt used for Zero-shot ICL with LLM

Figure 6 is the prompt used for ASAG dataset. The model is guided to assign grades while disregarding spelling errors and focusing on the content of the student's response. Figure 7 is the prompt used for Task 1 and Task 3-8 of the AES dataset, where the range of possible scores is task-specific and highlighted in blue, where `[[MIN]]` is the minimum score and `[[MAX]]` the maximum. Figure 8 presents the prompt for Task 2 of the AES dataset, which is the only task differing from the other eight tasks in its focus on grading according to *Writing Applications*. Throughout all figures, the orange text represents the model's initial response.

Instruction and Model answer

```
You are a teacher who assigns grades (between 1 and 7) to fourth grade students' answers to open-ended math questions. Do not discount for spelling or grammatical errors. Focus on what the student is trying to say with his or her answer. Indicate the grade after the prefix 'Score:'

Got it. Give me a question between <Q></Q> and an answer between <A></A>. Then ask me 'What is the score for the answer to the question?'. I will then give you the note after the prefix 'Score:'
```

Follow-up instruction

```
Question: <Q> [[Q]] </Q> Answer: <A> [[A]] </A> What is the score for the answer to the question?
```

Figure 6: Prompt used for Zero-shot ICL with LLM on the ASAG dataset. Translated from Spanish. the orange text represents a model's initial response.

Feature Name	Type of Feature
Exist a number in the answer	EF (Morphological)
Number of digits in the answer	EF (Morphological)
Number of numerical values in the answer	EF (Morphological)
The answer is composed of digits	EF (Morphological)
The answer is NaN (Not a Number)	EF (Morphological)
Proportion of digit characters in the answer	EF (Morphological)
Number of tokens in the answer	EF (Syntactic)
Number of tokens that do not contain numbers	EF (Syntactic)
Ratio of non-numeric tokens to the total number of tokens	EF (Syntactic)
Ratio of punctuation marks to the total number of tokens	EF (Syntactic)
Ratio of vowels to the total number of tokens	EF (Syntactic)
Length of the negation of the answer	EF (Syntactic)
Frequency of character 'x' in the answer	EF (Lexical)
Frequency of character 'y' in the answer	EF (Lexical)
Frequency of character 'g' in the answer	EF (Lexical)
Frequency of character 'h' in the answer	EF (Lexical)
Frequency of character 'j' in the answer	EF (Lexical)
Frequency of character 'k' in the answer	EF (Lexical)
Frequency of character 'w' in the answer	EF (Lexical)
Frequency of character 'ñ' in the answer	EF (Lexical)
Number of characters in the answer	EF (Structural)
Length of the longest number in the answer	EF (Structural)
Length of the longest sequence of repeated characters	EF (Structural)
Maximum number of consecutive vowels in a token	EF (Structural)
Maximum number of consecutive non-vowel characters in a token	EF (Structural)
Number of punctuation marks in the answer	EF (Punctuation)
Number of mathematical punctuation marks in the answer	EF (Punctuation)
Proportion of punctuation characters in the answer	EF (Punctuation)
Proportion of non-mathematical punctuation characters	EF (Punctuation)
Proportion of punctuation and digit characters in the answer	EF (Punctuation)
Proportion of non-digit and non-mathematical punctuation characters	EF (Punctuation)
Proportion of alphabetic characters that are vowels	EF (Phonological)
Proportion of vowels in the answer	EF (Phonological)
The answer shows a correct calculation of a quantity	NLLF
The answer does not show a correct calculation of a quantity	NLLF
The answer explains why a character is wrong	NLLF
The answer does not explain why a character is wrong	NLLF
The answer is free of conceptual errors	NLLF
The answer contains conceptual errors	NLLF
The answer shows a correct understanding of the question	NLLF
The answer does not show a correct understanding of the question	NLLF
The answer correctly indicates a quantity	NLLF
The answer does not correctly indicate a quantity	NLLF
The answer is written in a way that can be easily understood	NLLF
The answer is not written in a way that can be easily understood	NLLF
The answer is written clearly and concisely	NLLF
The answer is not written clearly and concisely	NLLF
The answer is correctly written in numerical format	NLLF
The answer is not correctly written in numerical format	NLLF
The answer is accompanied by an explanation	NLLF
The answer is not accompanied by an explanation	NLLF
The answer is complete and does not lack any relevant information	NLLF
The answer is incomplete or lacks relevant information	NLLF
The answer addresses the question	NLLF
The answer does not address the question	NLLF
The answer correctly identifies the value	NLLF
The answer does not correctly identify the value	NLLF

Table 6: Expert features (EF) and Natural Language Learned Features (NLLF) for the ASAG task Everything was translated from Spanish.

Feature Name	Code
Long-Word Ratio	RIX
Polysyllabic Grade Level	SMOG
Complex Word Grade Level	GF
Familiar Word Difficulty	DC
Number of sentences	S
Number of adjective words	JJ
Number of unique words	UW
Number of preposition / subordinating - conjunction words	IN
Number of long words	LW
Number of determiner words	DT
Number of difficult words	DW
Number of complex words	CW
Number of noun words	NN
Number of commas	CO
Number of wordtype	WT
Number of non-basic words	NBW
Number of words	W
Number of characters	CH
Number of adverb words	RB

Table 7: Linguistic Features from Wang et al. (2023) for the AES task.

Method	SF	Text	EF	NLLF	EF + NLLF
ULRA	-	.6423	-	-	-
LR	✗	-	.5712	.6041	.6227
	✓	-	.5707	.6035	.6193

Table 8: Results of the Logistic Regression model using the scores of ULRA as a target during the weakly supervised learning. SF is Signal Filtering.

Instruction and Model answer

```
You are a teacher who assigns grades (between [[MIN]] and [[MAX]]) to essays from students ranging in grade levels from Grade 7 to Grade 10. You will help me break down the 'assign grade to student essay' task. To do this, I will give you a sample essay along with the assignment. Indicates the score after the prefix 'Score:'.

Got it. Give me a question between <A></A> and an essay between <E></E>. Then ask me 'What is the score for the essay?'. I will then give you the score after the prefix 'Score:'.
```

Follow-up instruction

```
Assignment: <A> [[A]] </A> Essay: <E> [[E]] </E> What is the score for the essay?
```

Figure 7: Prompt used for Zero-shot ICL with LLM on the Task 1 and Tasks 3 to 8 of the AES dataset. The blue text highlights the range of values specific to each task, while the orange text represents a model's initial response.

Instruction and Model answer

```
You are a teacher who assigns grades (between 1 and 6) to essays from students ranging in grade levels from Grade 7 to Grade 10. You will help me break down the 'assign grade to student essay according to Writing Applications' task. To do this, I will give you a sample essay along with the assignment. Indicates the score after the prefix 'Score:'.

Got it. Give me a question between <A></A> and an essay between <E></E>. Then ask me 'According to Writing Applications, what is the score for the essay?'. I will then give you the score after the prefix 'Score:'.
```

Follow-up instruction

```
Assignment: <A> [[A]] </A> Essay: <E> [[E]] </E> According to Writing Applications, what is the score for the essay?
```

Figure 8: Prompt used for Zero-shot ICL with LLM on the Task 2 of the AES dataset. The orange text represents a model's initial response.

Feature Name	Coef.	Std. err.	[0.025	0.975]
Intercept	4.65	0.01	4.64	4.67
The answer is correctly written in numerical format	0.00	0.00	0.00	0.00
The answer is written in a way that can be easily understood	0.50	0.01	0.47	0.52
The answer shows a correct calculation of a quantity	0.00	0.00	0.00	0.00
The answer correctly identifies the value	0.00	0.00	0.00	0.00
The answer shows a correct understanding of the question	0.00	0.00	0.00	0.00
The answer explains why a character is wrong	0.12	0.01	0.10	0.13
The answer is accompanied by an explanation	0.00	0.01	0.00	0.03
Exist a number in the answer	0.27	0.01	0.24	0.30
Frequency of character 'g' in the answer	0.00	0.00	0.00	0.00
Frequency of character 'h' in the answer	0.00	0.00	0.00	0.00
Frequency of character 'k' in the answer	0.00	0.00	0.00	0.00
Frequency of character 'w' in the answer	0.00	0.00	0.00	0.00
Frequency of character 'x' in the answer	0.03	0.01	0.02	0.04
Frequency of character 'y' in the answer	0.11	0.02	0.07	0.14
Number of characters of the answer	0.14	0.12	0.00	0.35
Number of tokens in answer	0.00	0.00	0.00	0.00
Length of the negation of the answer	0.00	0.06	0.00	0.21
Length of the longest number in the answer	0.11	0.01	0.08	0.13
Maximum number of consecutive non- vowel characters in a token	0.15	0.01	0.12	0.17
Number of digits in the answer	0.00	0.00	0.00	0.00
Number of mathematical punctuation marks in the answer	0.00	0.01	0.00	0.02
Number of tokens that do not contain numbers	0.35	0.18	0.00	0.59
Number of numerical values in the answer	0.00	0.00	0.00	0.00
Number of tokens in the answer	0.00	0.00	0.00	0.00
Proportion of alphabetic characters that are vowels	0.43	0.04	0.35	0.52
Proportion of punctuation characters in the answer	0.00	0.00	0.00	0.00
Proportion of punctuation and non- vowel characters in the answer	0.14	0.01	0.10	0.16
Proportion of vowels in the answer	0.21	0.05	0.10	0.30
Ratio of non-numeric tokens to the total number of tokens	0.00	0.02	0.00	0.05
Ratio of punctuation marks to the total number of tokens	0.00	0.00	0.00	0.00

Table 9: Coefficients of the Linear Regression with Signal Filtering using EF+NLLF features in the ASAG dataset. $[0.025, 0.975]$ refers to the 95% confidence interval of the coefficient.

F Human validation of the NLLFs

F.1 NLLFG Classifiers

Here we analyze how accurate were the NLLF generated by the BERT-like model, and also the weak labels by the LLM. We took 190 examples from the validation set used to train the NLLFG of the ASAP task, and asked an expert to manually label them regarding the labels of a BSQ. More precisely, we manually annotated 10 examples sampled uniformly per BSQ having non-zero weights in the linear regressions (approximately 2-3 BSQs per essay set) across 8 essay sets. We compare the labeling of the expert with the outputs of the NLLFG and LLM models, using classical classification metrics such as precision, recall and F1-score.

The results for both the models are available in Table 10. The LLM obtained a better F1-score than the smaller transformer model, which was expected. It is interesting to note that the accuracy of the NLLFG model is 0.78, close to the ones of the LLM (0.86). The macro F1-scores are more divergent as the LLM reaches 0.84 and the NLLFG 0.74, which is still better than random.

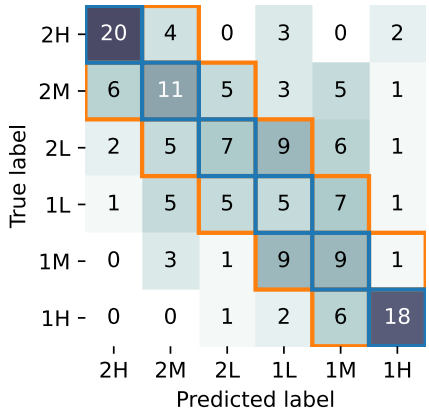
Model	Label	Prec.	Rec.	F1	Acc.
Mixtral	Yes	91	88	89	86
	No	76	82	79	
NLLFG	Yes	90	78	84	78
	No	57	76	65	

Table 10: Performance of NLLFG and Mixtral on a manually annotated set of 190 examples. The dataset consists of 10 uniformly selected examples per BSQ (approximately 2-3 BSQs per essay set) across 8 essay sets.

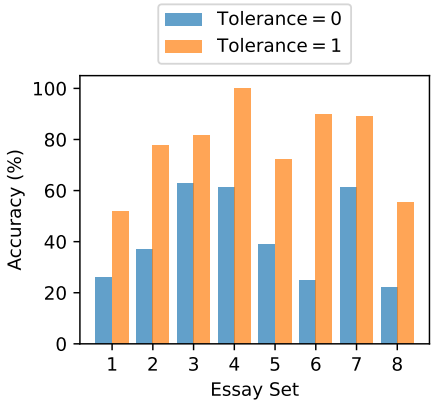
F.2 NLLFs Before the Linear Regression

We designed another experiment to assess the reliability of the NLLF with respect to human annotation, showing pairs of examples to a human, and asking which should have the highest value in NLLF and what is the distance in values between the examples of the pair. As the NLLF are normalized before the linear regression, hence each score depends on the whole group and becomes relatives to the other examples (the best has a highly positive score and the worst has a highly negative score).

Pairs of examples with various distances in-



(a) Global Confusion Matrix



(b) Accuracies with tolerance 0 and 1

Figure 9: Metrics between the human annotation and the real values of the NLLFs, for the AES task.

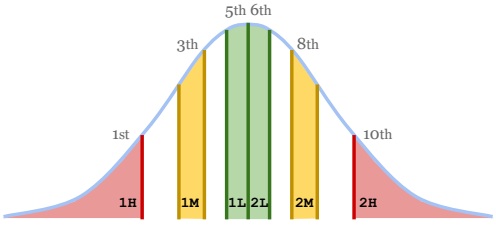


Figure 10: Examples were picked from bins with **high**, **medium** and **low** distance between each others. For a pair of examples, the annotator has to find which example has the highest value, and what is the distance between the examples.

Method	Weak Signal	Signal Filt.	Text	EF	NLLF	EF + NLLF
Length	None	-	0.0015	-	-	-
Jaccard Sim.	None	-	-0.1335	-	-	-
Jaccard Sim.	None	-	0.3170	-	-	-
ULRA	LF	-	0.4562	-	-	-
	EF	-	0.3713	-	-	-
	EF+LF	-	0.3902	-	-	-
Z-score	None	-	0.4346	-	-	-
LLM	None	-	**0.5629**	-	-	-
LLM-CoT	None	-	0.4631	-	-	-
Linear Regression	Z-score	✗	-	0.4472	0.3627	0.4167
	LLM-based signal	✗	-	0.4212	0.2984	0.3772
	Z-score	✓	-	0.4471	0.3435	**0.4915**
	LLM-based signal	✓	-	0.3682	0.2925	0.4115
BERT	Z-score	✗	0.3965	-	-	-
	LLM-based signal	✗	0.3867	-	-	-
	Z-score	✓	0.2451	-	-	-
	LLM-based signal	✓	0.3848	-	-	-
Human	None	-	**0.7403**	-	-	-

Table 11: Results on ASAG using the QWK

between the examples were randomly selected regarding their places in the distributions: pairs from the first and last deciles of the distribution, pairs from the 3rd and 8th, and pairs from the 5th and 6th. We ask a human to tell for each pair, which example is the highest in the distribution, and how large is the distance between them. It gave us a classification problem with 6 ordinal classes: First-High (1H), First-Medium (1M), First-Low (1L), Second-Low (2L), Second-Medium (2M), Second-High (2H).

We focused on the 19 BSQs having non-zero weights in the linear regressions, and randomly selected 3 examples of High, Medium and Low distances between the pairs, which gave us a total of 171 pairs to annotate coming from 6 classes. Figure 10 shows the bins of the examples from the different categories.

The results overall are shown in Figure 9, with the confusion matrix and the We report a Krippendorff (2013)'s alpha of 0.63, an Accuracy of 0.43 (random is 0.17) and an accuracy with a tolerance of 1 (Gaudette and Japkowicz, 2009) of 0.77 (random is 0.44). This shows that human rank the examples in an order similar to the ones of the NLLF values 77% of the time using a tolerance of 1 in the ordinal classification.

G Others

Table 11 shows the results on the ASAG dataset using QWK. The results are very similar: LLM is better than our method, which is itself better than ULRA.