Automated L2 Proficiency Scoring: Weak Supervision, Large Language Models, and Statistical Guarantees

Aitor Arronte Alvarez
University of Hawai'i at Mānoa
Honolulu, HI, USA
arronte@hawaii.edu

Naiyi Xie Fincham
University of Hawai'i at Mānoa
Honolulu, HI, USA
naiyixf@hawaii.edu

Abstract

Weakly supervised learning (WSL) is a machine learning approach used when labeled data is scarce or expensive to obtain. In such scenarios, models are trained using weaker supervision sources instead of human-annotated data. However, these sources are often noisy and may introduce unquantified biases during training. This issue is particularly pronounced in automated scoring (AS) of second language (L2) learner output, where high variability and limited generalizability pose significant challenges. In this paper, we investigate the analytical scoring of L2 learner responses under weak and semi-supervised learning conditions, leveraging Prediction-Powered Inference (PPI) to provide statistical guarantees on score validity. We compare two approaches: (1) synthetic scoring using large language models (LLMs), and (2) a semi-supervised setting in which a machine learning model, trained on a small gold-standard set, generates predictions for a larger unlabeled corpus. In both cases, PPI is applied to construct valid confidence intervals for assessing the reliability of the predicted scores. Our analysis, based on a dataset of L2 learner conversations with an AI agent, shows that PPI is highly informative for evaluating the quality of weakly annotated data. Moreover, we demonstrate that PPI can increase the effective sample size by over 150% relative to the original human-scored subset, enabling more robust inference in educational assessment settings where labeled data is scarce.

1 Introduction

Recent advances in Natural Language Processing (NLP) have enabled the development of intelligent conversational agents for language learning and teaching that are capable of producing human-like language. In the context of computer-assisted language learning (CALL), research-driven, dialogue-based systems, such as task-specific conversational agents designed to support second language (L2) acquisition, have shown promising results in fostering vocabulary and grammatical development, while also promoting self-directed learning through repeated, skills-focused practice (Bibauw et al., 2019; Tyen et al., 2022; Glandorf et al., 2025).

These technological developments have significantly enhanced the ability of dialogue-based CALL systems to guide and sustain human-like conversational interactions, aligning them with established proficiency guidelines and pedagogical principles. As a result, they offer structured, reliable, and personalized L2 practice beyond the classroom. This shift underscores the need for scalable, efficient, and statistically valid assessment methods capable of supporting such learning environments.

Automated scoring (AS) of language output, such as written essays (Shermis and Burstein, 2013), short texts (Burrows et al., 2015), spoken dialogues (Litman et al., 2018), and text-based conversations (Ramanarayanan et al., 2019; Yuwono et al., 2019), is a mature field of research that emerged during the 1960's (Page, 1968) and has accelerated its development over the past two decades (Shermis and Burstein, 2003; Xi, 2010; Ke and Ng, 2019) as NLP methods have evolved significantly. However, AS methods rely on large quantities of high-quality manually annotated data to train models, which requires significant human resources and time.

To overcome the difficulties and challenges of data annotation in NLP, Weakly-supervised learning (WSL) emerged as an alternative framework (Huang et al., 2014), leveraging weaker sources and methods to obtain synthetic labels from textual data. Many of the strengths of WSL depend on the availability of high-quality validation data (Zhu et al., 2023), which in L2 assessment, is not always possible. Assessing L2 output for learning requires not only knowledge of the target language but also the ability to evaluate a learner's interlanguage based on established proficiency guidelines,

Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications, pages 384–397

making it an even more time-consuming task.

With the development of large language models (LLMs) and their advanced language understanding capabilities, researchers have begun to utilize them in data annotation tasks (Goel et al., 2023; Tan et al., 2024b). In L2 assessment in particular, GPT-4 has shown to produce holistic scores that are highly correlated to human evaluation in written essays and have moderate to high inter-rater reliability (Tate et al., 2024). Furthermore, experiments showed that GPT-4 is capable of performing analytic scoring of L2 texts given holistic scores (Banno et al., 2024), however, no ground truth set was available in this study. While state-of-the-art LLMs such as GPT-4 have shown human-like language capabilities that allow them to produce annotations that are highly correlated with expert ones, it is unclear how biased those annotations are. In addition, no statistical guarantees on the validity of the synthetic data are used in the literature.

In this paper, we investigate whether state-of-the-art large language models (LLMs) can be used to generate high-quality synthetic scores of lexical complexity and grammatical accuracy from students' text-based conversational responses based on the Common European Framework of Reference (CEFR) framework (Council of Europe, 2001). These synthetic scores, along with a small set of human-annotated gold-standard data, are used to train machine learning models under two different settings: a weakly supervised learning (WSL) approach that relies on LLM-generated labels, and a semi-supervised method in which a model trained on the gold-standard set produces predictions for a larger unlabeled corpus. In both settings, our goal is to increase the effective sample size and enable valid inference. To this end, we apply Prediction-Powered Inference (PPI) to provide statistical guarantees on the resulting predictions, ensuring that the use of synthetic scores does not compromise the validity of the conclusions.

Experimental results indicate that the proposed method increases the effective sample size by over 150% and yields a relative gain in accuracy, both compared to using only the gold-standard human-annotated data in a semi-supervised setting. In contrast, treating LLM-generated scores as if they were human-annotated can lead to inaccurate estimates and yield more modest improvements under a WSL framework. The proposed approach helps mitigate some of the limitations associated with weaker supervision sources in NLP, particularly in scenarios where predictions inform decisions with significant consequences, such as in educational assessment.

We also address challenges associated with using LLMs as data annotators in NLP tasks, especially the uncertainty inherent in their outputs. Our findings show that applying a statistically valid method such as PPI can not only improve reliability and provide bias corrected estimates, but also quantify the uncertainty of predictions on unlabeled data, thereby offering a more trustworthy framework for leveraging synthetic annotations and scores.

The main contributions of this paper are:

- We integrate Prediction-Powered Inference into a new framework for semi-supervised and weakly supervised learning, providing statistical guarantees for predictions on datasets with small labeled and large unlabeled subsets.
- Unlike standard semi- and weakly supervised learning paradigms, the proposed framework samples and selects synthetic data based on valid statistical conditions, imposing a data quality requirement relative to a gold standard set.
- This approach, in the semi-supervised setting, produces a relative sample size gain of up to 157%, resulting in an accuracy increase of 23.2%.

2 Background

2.1 Automated L2 scoring methods

Over the past decades, computer-aided automatic text analysis has become increasingly prevalent in measuring L2 lexical and speaking proficiency (Crossley et al., 2011, 2014). More recently, deep learning approaches have achieved performance close to that of human raters in holistic scoring tasks (Alikaniotis et al., 2016), and Transformer-based models have even surpassed human inter-annotator agreement levels (Rodriguez et al., 2019). Large language models (LLMs) such as GPT-3 have also shown promise in supporting automatic scoring, as demonstrated by their application to 12,100 essays from the ETS Corpus of Non-Native Written English (Mizumoto and Eguchi, 2023).

Further advancements have been observed with GPT-4. Studies indicate that, when provided with calibration examples, GPT-4 can reliably rate short essay responses (Yancey et al., 2023), assess discourse coherence at a level comparable to expert

raters (Naismith et al., 2023), and generate analytical scores aligned with the CEFR proficiency framework (Banno et al., 2024).

While NLP-based automated methods have historically demonstrated the ability to assess specific linguistic features and functions, human raters tend to outperform them in evaluating higher-level discourse elements such as ideas, content, and organization (Enright and Quinlan, 2010). This divergence suggests that language models may exhibit a different type of bias compared to human raters, particularly in tasks requiring inferential judgment.

2.2 Weaker sources of supervision

Weakly-supervised learning (WSL) has become a practical machine learning paradigm to address the issue of label scarcity in NLP. The major bottleneck for deploying machine learning models has been the lack of access to large, high-quality training datasets. Producing manual annotations of text data is a labor-intensive and time-consuming task. To reduce such efforts, WSL approaches have been proposed to offer a larger pool of weaker supervision sources to label and annotate data (Ren et al., 2020; Zhang et al., 2021). Such sources often rely on heuristics, knowledge bases, crowd sourcing, labeling functions, or pre-trained models instead of expert manual annotations (Ratner et al., 2017). However, WSL methods also present challenges due to the degree of noise that the generated labels contain (Zhu et al., 2023).

More recently, a prompting-based method was proposed to integrate LLMs into weak supervision frameworks (Smith et al., 2024), yielding accuracy gains on the general-purpose WRENCH weak supervision benchmark. However, the effectiveness of this approach in more specialized domains, such as the analytical scoring of student responses, remains uncertain. Moreover, the study does not address potential biases present in the training data, nor does it evaluate how such biases may affect the resulting estimates. It also remains unclear how a semi-supervised method (Søgaard, 2022) would perform in comparison to this weakly supervised approach, particularly in settings where bias is limited to the human annotations and model predictions, without introducing additional external sources of error.

To address this gap, our study compares both approaches, LLM-driven weak supervision and a semi-supervised method using a small gold-standard dataset, to investigate their effectiveness in analytical scoring tasks. We leverage PPI in both cases to provide statistical guarantees on the resulting predictions and to evaluate the reliability and calibration of the scores derived from each approach.

3 Method

We are interested in developing a framework that can be used to train machine learning models when only a small labeled dataset and a large corpus of unlabeled data are available. To leverage the unlabeled data, synthetic scores are obtained using a machine learning model, but instead of treating those scores as gold standard, a provably valid statistical method is used to assess the biases contained in the scores, so that estimates can be rectified.

The proposed framework is evaluated in two settings. In the weakly supervised learning (WSL) scenario, a state-of-the-art LLM is used to generate synthetic scores from text-based inputs, which are then used to train a traditional machine learning model. In the semi-supervised setting, the ML model is trained on a small set of gold-standard annotations and used to predict scores for a larger unlabeled set. In both cases, a debiasing protocol based on Prediction-Powered Inference (PPI) (Angelopoulos et al., 2023) is applied to estimate prediction errors and provide statistical confidence measures for the resulting scores.

Although LLMs have shown strong zero-shot generative and reasoning capabilities (Kojima et al., 2022), they still produce hallucinations (Gunjal et al., 2024), unreliable outputs (Sclar et al., 2023), and exhibit demographic biases (Chiang and Lee, 2023), making them unreliable for providing immediate scores to students. For those reasons we use machine learning models that can be trained on a set of textual features and a combination of human and weaker scoring sources to estimate a proficiency score with a given confidence (see details on models and features in subsection 4.2).

While NLP tasks, particularly those in the social sciences, have used less reliable LLM annotations in downstream tasks that require inferences to be statistically valid to draw reliable conclusions (Gligorić et al., 2024), the approach presented in this paper leverages such statistical validity to determine the reliability of weaker data sources to train machine learning models.

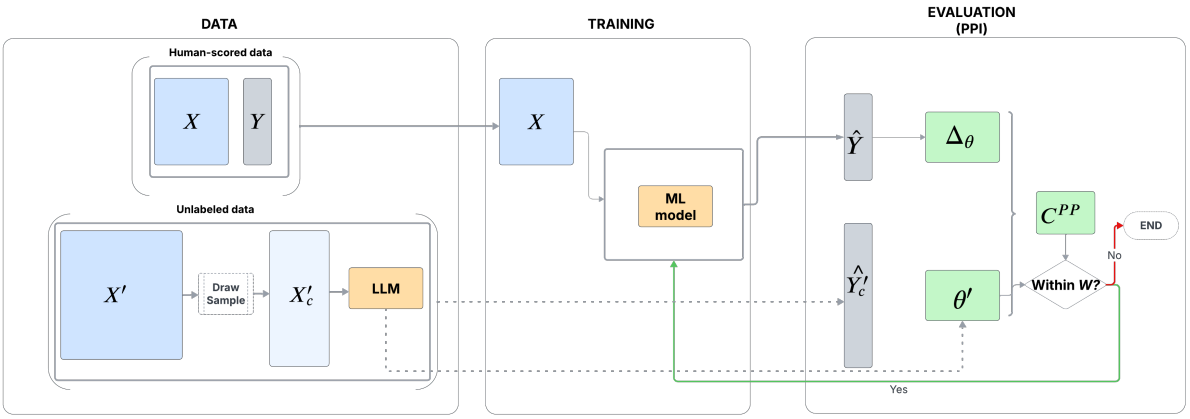

Figure 1: Outline of the process for scoring L2 conversation responses in a WSL setting using PPI. X and Y are the features and human-annotated scores, and X'_c are a subset of the textual features sampled from the unlabelled dataset X' to be scored by an LLM, obtaining $\hat{Y}'_c$ scores and θ' verbalized confidence on the scores. If the width of C^{PP} remains within W X'_c will be added to the training process to further optimize the training of the ML model.

3.1 Statistical guarantees: Prediction-powered inference

Prediction-Powered Inference (PPI) is a statistical protocol that combines predictions made on less reliable unlabeled data with those made on a gold-standard dataset to obtain a confidence interval (CI) that is provably valid (Angelopoulos et al., 2023). Instead of using machine learning models to determine the validity of an unlabeled dataset on a case-by-case basis, PPI provides model-free estimates that are statistically valid, leveraging the information contained in the predictions.

The goal of PPI is to estimate a quantity of interest θ^*, such as the population mean. To estimate θ^* we have access to a set of gold-standard data with human-annotated responses Y and features X such that $(X, Y) = (X_1, Y_1), ..., (X_n, Y_n)$, and a much larger set of unlabeled data $(X', Y') = (X'_1, Y'_1), ..., (X'_N, Y'_N)$ where Y' is not directly observable, and $N \gg n$. For both datasets predictions are obtained using a machine learning model $f(\cdot)$, represented by $f(X)$ and $f(X')$. In PPI, the predictions made on the unlabeled data are not treated as gold-standard such as in the imputation case. Instead, PPI uses the gold-standard set to quantify and correct for the errors made by the model on the unlabeled set.

The three-step process that constitutes PPI can be summarized as follows:

1. Select the quantity of interest θ^*, such as the mean outcome $\mathbb{E}(Y_i)$.

2. Compute the estimate θ' and a rectifier Δ_θ, where θ' is computed on the unlabeled data $(X', \hat{Y}')$ such that $\theta' = \frac{1}{N}\sum_{i=1}^{N} f(X'_i)$, and $\Delta_\theta = \frac{1}{n}\sum_{i=1}^{n}(f(X_i) - Y_i)$. If $f(X_i)$ perfectly matches Y, then $\Delta_\theta = 0$.

3. Construct a confidence interval C^{PP} for θ^*.

To construct C^{PP} we need to obtain the prediction-powered estimate $\hat{\theta}^{PP}$ that corrects for the bias on θ' due to prediction errors:

$$\hat{\theta}^{PP} = \frac{1}{N}\sum_{i=1}^{N} f(X'_i) - \frac{1}{n}\sum_{i=1}^{n}(f(X_i) - Y_i) \quad (1)$$

and then the prediction-powered confidence set is obtained such that

$$C^{PP} = (\hat{\theta}^{PP} \pm w(\alpha)) \quad (2)$$

where $w(\alpha)$ is a constant that depends on the confidence level α (derivations could be found in Angelopoulos et al. (2023)).

PPI has been used for the pairwise ranking of models (Boyeau et al., 2024), for comparing the performance of LLMs (Chatzi et al., 2024), for evaluating retrieval augmented generation (RAG) systems (Saad-Falcon et al., 2024), and some of its variants for producing confident conclusions from LLMs annotations (Gligorić et al., 2024). The approach presented in this article differs from the previous ones. In the general PPI setting, a trained model is used to produce predictions on both sets, and PPI is used to debias the predictions made on the unlabeled data. In the proposed framework, we do not have access to a trained model, and the

training is done iteratively and sequentially using PPI as a guarantee of statistical validity, making decisions on what unlabeled data to include in the training process based on a statistical measure.

3.2 Using LLMs in weak supervision with statistical guarantees

From a dataset of conversational responses $\mathcal{X}$, we divide it into $|X|$ responses to be scored by human annotators and $|X'|$ responses to be scored by an LLM, obtaining Y and Y' scores respectively; where $|X| = n$ and $|X'| = N$, and $N \gg n$. We assume that there are biases in $\hat{Y}'$ associated with the scoring errors made by the LLM, and use PPI to debiase them, resulting in biased-corrected estimates.

To obtain Y and Y', the same rubric was used for human and LLM scorers, in an attempt to maintain as much parity as possible between the two scoring sources and to avoid additional biases.

The rubric used two dimensions of language proficiency as expressed in the CEFR framework (Council of Europe, 2001), namely vocabulary range and grammatical accuracy for B1 and B2 levels. Scores ranged from 1-3 for vocabulary range and 1-4 for grammatical accuracy. For human scoring, two annotators with extensive experience in L2 proficiency scoring were recruited. To obtain a single score, annotations were conducted collaboratively, and if consensus was not reached a third annotator was used to resolve the disagreement (Fort, 2016) (see Appendix A for details on the rubric). GPT-4o and GPT-4o-mini were the models of choice to produce synthetic scores Y' given X' and the rubric. A zero-shot prompting approach was used (see Appendix C for details on the prompts) and additionally, the models were prompted to provide a measure of *verbalized confidence* in the form of a probability value to assess the correctness of the score, as presented in Tian et al. (2023).

The following steps describe the WSL approach with an LLM as a *weak* scorer and with PPI guarantees: 1) taking as input the entire set of high-quality human-labeled scores, a machine learning model $f(\cdot)$ is trained such that, after training on the gold-standard set X is completed, we obtain $\hat{\theta}^{PP}$ and C^{PP} using the verbalized confidence of the LLM θ' on a small sample of size c X'_c and the predictions made by the ML model $\hat{Y}$; 2) the width of C^{PP} is computed in an evaluation step, such that $C^{PP}_{upper} - C^{PP}_{lower} \leq W$ and W is a width threshold chosen beforehand; 3) if the width is not greater than W and the prediction-powered corrected mean accuracy is not less than the one computed with the gold-standard set, the sample X'_c is added to the training process and the model is trained on $X \oplus X'_c$ until either the C^{PP} condition on W is no longer met or the accuracy decreases. Figure 1 outlines the process in a block diagram.

3.3 Leveraging PPI in semi-supervised learning

Similar to the WSL approach described in Subsection 3.2, the proposed semi-supervised method uses PPI to establish statistical guarantees on predictions made for the unlabeled data X'. However, instead of relying on an LLM as a scorer, this method employs a machine learning model to generate predictions, which are then reused for fine-tuning under the same width and accuracy gain conditions defined in the WSL setting.

The method works as follows. First, the ML model is trained only on the human-scored set, the ground-truth data X. In an evaluation step, a sample of size c is randomly drawn from the entire unlabeled set X' to compute $\hat{Y}'_c = f(X'_c)$ and obtain $\hat{\theta}^{PP}$ and C^{PP}. If the width of C^{PP} does not exceed a threshold W, i.e., $C^{PP}_{upper} - C^{PP}_{lower} \leq W$, and $\hat{\theta}^{PP}$ is greater than the mean accuracy of the predictions made using the human-scored data, then a new sample is drawn and training continues until this condition is no longer satisfied (Figure 2 summarizes the steps involved in this process).

As we can see, PPI is used to estimate the validity of the inferences made on the unlabeled set through an iterative process that draws samples of size c to test whether the predictions on X' maintain the width of C^{PP} within the threshold W. A wider width would denote greater uncertainty, indicating that the predictions made on the unlabeled data are less reliable and potentially more biased. In contrast, a narrower width suggests higher precision and lower variance.

In this sense, C^{PP} serves as a valid estimate for assessing the quality of an unlabeled dataset given a high-quality labeled one. It also provides a basis for estimating the effective sample size needed to obtain reliable predictions when leveraging unlabeled data, especially when considering the associated accuracy gain.

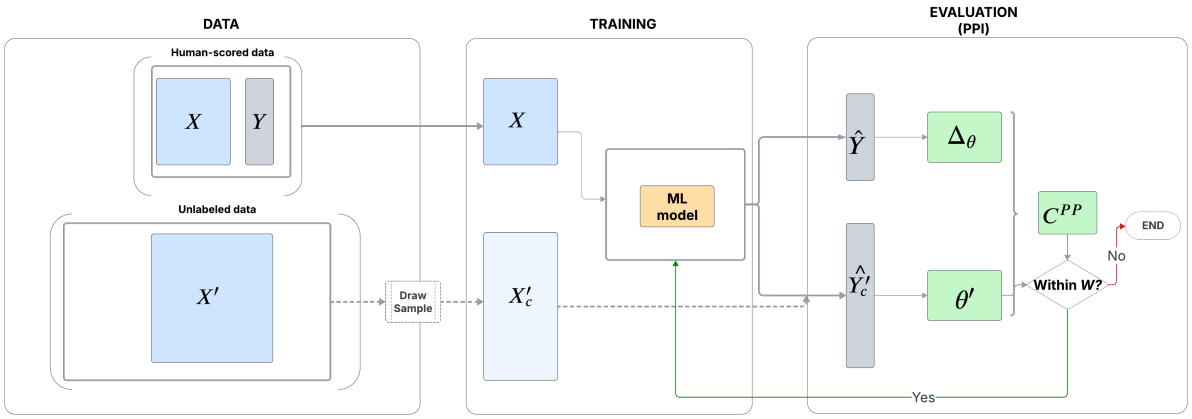

Figure 2: Differently from the process outlined in Figure 1, in the semi-supervised setting both the scores $\hat{Y}'_c$ and the probability value θ' are obtained directly from the ML model. No external scoring source is used.

4 Experiments

We conduct several experiments to evaluate the effectiveness of our approach. The goal is to assess the overall methodology in both weakly and semi-supervised learning settings, aiming to measure the quality of synthetically generated scores derived from a smaller set of high-quality, human-annotated data. Given the constraints of this study, namely the limited availability of high-quality human-labeled samples, we use machine learning models that are well-suited to this low-resource setting and that have shown to perform effectively on features that can be represented as tabular data (Shwartz-Ziv and Armon, 2022). We make code available [1].

4.1 Data

Data was collected from text-based conversation practice sessions completed by intermediate level (B1-B2 levels in CEFR) English language learners (ELLs) and an AI agent (Fincham and Alvarez, 2024) over a 3-month period. A total 121 students from 3 sessions of an undergraduate course focused on English speaking participated in the project and generated 1721 practice sessions. The average number of turns per session produced by students was 8.9.

To train the models, 590 sessions were manually scored following a rubric based on the CEFR framework (Council of Europe, 2001) on vocabulary range and grammatical accuracy (see section 3.2). Out of the 590 sessions, 445 were used for training and 145 for evaluation. The remaining 1131 sessions were either scored by an LLM or by the model of choice in the semi-supervised experiment. Inter-annotator agreement between human and LLM raters reached moderate levels, with $\kappa = 0.45$ for vocabulary range and $\kappa = 0.4$ for grammatical accuracy.

4.2 Models and features

From the students' conversations, 9 lexical and syntactical features were automatically extracted, many of which have shown to be highly correlated with linguistic proficiency descriptors based on the CEFR framework (Banno et al., 2024). Those are: lexical density, unique noun chunks, number of unique words, number of unique difficult words, Flesch Kincaid readability score (Thomas et al., 1975), sentence length mean and standard deviation, and dependency distance mean and standard deviation.

Two tree-based boosting models, XGBoost (Chen and Guestrin, 2016) and LightGBM (Ke et al., 2017) were used in the weak and semi-supervised training regimes and used as predictor ML models. Results were compared to the baseline scores obtained directly from the two LLMs, GPT-4o and GPT-4o-mini.

4.3 Evaluation

The quantity of interest chosen for this study was the mean accuracy. As described in subsections 3.2 and 3.3, width and accuracy gain are the measures that determine the stopping condition during training and evaluation, and overall, to determine the quality of the inference on the unlabeled data. In addition, coverage is used to evaluate how many times the true value θ^* falls within the estimated

[1] https://github.com/aitor-alvarez/Automated-L2-Proficiency-Scoring

interval with a given confidence level. The confidence level for this study was set to 0.9 for $\alpha = 0.1$, which is the same level that PPI guarantees asymptotically (Angelopoulos et al., 2023). If coverage does not meet the established $\alpha-$level, this would indicate that predictions are extremely biased, not normally distributed, or that the proportion or quality of the labeled/unlabeled data is not balanced and therefore the variance estimate may be unstable.

We estimate the effective sample size by calculating the maximum number of synthetic scores used in relation to the labeled set for the following inequality to hold $C^{PP}_{upper} - C^{PP}_{lower} \leq W$ and for the mean accuracy to improve when comparing it to the results obtained with the gold-standard set alone. The width threshold was set at $W = 0.2$ and tested in the two experimental learning settings (weakly and semi-supervised) for each of the models. This width threshold indicates that we are willing to accept a CI with a maximum of 20% range in the mean accuracy estimate C^{PP}. Samples of size 100 were added at each iteration to determine the width, coverage, and effective sample size for both conditions.

5 Results

Table 1 presents the experimental results. The semi-supervised approach yields the highest accuracy gains by using PPI to combine human-annotated data with model-generated scores, selecting only samples within the PPI confidence interval that improve baseline accuracy. For vocabulary range, the increase reaches 23.2% and for grammatical accuracy 21.5%, with a total effective sample size of 700. Width sizes remain relatively low, 0.13 for vocabulary range and at around 0.148 for grammatical accuracy. Coverage in this setting reaches 97%, demonstrating the validity of this approach.

The weakly supervised learning (WSL) protocol, on the other hand, yields more modest accuracy gains when combining gold-standard data with weakly scored data. In this setting, accuracy improvements range from 8.1% to 8.4% in vocabulary range and reach 7.5% in grammatical accuracy, using either boosting model with GPT-4o as the annotator source and an effective sample size of 200. When GPT-4o-mini is used as the annotator model, accuracy gain decreases to 4.2–3.4% in vocabulary range and 3.1% in grammatical accuracy, with an effective sample size of 100. Overall, the WSL setting shows a coverage slightly above 90% (91%), indicating an acceptable validity of this approach when using LLMs as weak scorers (see Appendix B for the accuracy on the gold-standard set only).

The LLM-only approach yields modest accuracy gains, with GPT-4o achieving improvements of 1.8% in vocabulary range and 1.5% in grammatical accuracy. GPT-4o-mini shows smaller gains of 0.6% and 0.4%, respectively, under the same setting, with an effective sample size of 100 in both cases. However, despite these gains, the coverage remains below 90%, failing to meet the required validity threshold. Further analysis reveals that both LLMs exhibit overconfidence in their probability estimates. On average, GPT-4o assigns 80% confidence to incorrect predictions in vocabulary range and 75% in grammatical accuracy. GPT-4o-mini shows similar patterns, with 78% confidence in vocabulary range and 71% in grammatical accuracy for its incorrect predictions.

In summary, the results indicate that the method presented in this study, when applied in a semi-supervised setting, results in a dataset that is 157% larger than the original. In contrast, the sample size gain is significantly reduced, down to 22%, when the method is used in a WSL setting with LLMs as scorers. Moreover, the naive approach of directly using LLM responses as gold-standard predictions fails to produce valid results.

6 Discussion

In this study, we have presented an approach to integrate Prediction-Powered Inference (PPI) in semi- and weakly-supervised settings when gold-standard data is scarce or difficult to obtain. By using PPI, we have shown that gold-standard with less reliable data can be combined to obtain increases in predictive accuracy while maintaining the validity of the results. This is particularly important in the context of this study, where student-produced output, namely conversational responses obtained from student interactions with an AI tutor, requires assessment at scale that can provide valid feedback to learners.

As previous studies have demonstrated (Tate et al., 2024; Tan et al., 2024b,a), LLM outputs show moderate to strong agreement with human judgments. In our study, we observe a moderate level of agreement between human raters and LLM-based scores, as previously reported. However, this level of agreement is insufficient for treating LLM scores as gold-standard, as it does not yield valid

Setting	Task	Model	Sample Size	Width	Coverage	Acc. Gain
Semi	Vocab. range	XGBoost	700	0.13	97%	23.2%
Semi	Gram. accur.	XGBoost	700	0.145	97%	21.5%
Semi	Vocab. range	LightGBM	700	0.133	97%	22.1%
Semi	Gram. accur.	LightGBM	700	0.148	97%	19.7%
WSL	Vocab. range	XGBoost + 4o	200	0.136	91 %	8.4%
WSL	Gram. accur.	XGBoost + 4o	200	0.144	91 %	7.5%
WSL	Vocab. range	XGBoost + 4o-mini	100	0.16	91 %	4.2%
WSL	Gram. accur.	XGBoost + 4o-mini	100	0.171	91 %	3%
WSL	Vocab. range	LightGBM+ 4o	200	0.14	91 %	8.1%
WSL	Gram. accur.	LightGBM+ 4o	200	0.145	91 %	7.5%
WSL	Vocab. range	LightGBM+ 4o-mini	100	0.177	91 %	3.4%
WSL	Gram. accur.	LightGBM+ 4o-mini	100	0.179	91 %	3.1%
LLM only	Vocab. range	GPT-4o	100	0.152	77%	1.8%
LLM only	Gram. accur.	GPT-4o	100	0.156	77%	1.5%
LLM only	Vocab. range	GPT-4o-mini	100	0.181	68%	0.6%
LLM only	Gram. accur.	GPT-4o-mini	100	0.184	68%	0.4%

Table 1: Performance metrics by setting, task, and model employed to obtain predictions and to generate synthetic scores. Sample size indicates the maximum number of unlabeled samples used to reach the highest accuracy and within the maximum width allowed. Acc. Gain is the maximum gain in accuracy compared to the gold-standard (human annotated only) approach (see Appendix B for the accuracy on the gold-standard set only).

statistical conclusions.

When LLM outputs are instead used within a weak supervision framework, acknowledged as biased but corrected through prediction-powered inference (PPI), they lead to slight improvements compared to relying solely on gold-standard data. Nonetheless, LLMs exhibit overconfidence in their incorrect assessments, indicating a poor understanding of uncertainty, a concern also noted in a recent work (Pawitan and Holmes, 2025).

In contrast, we find that a well-calibrated machine learning model, when used in a semi-supervised setting alongside PPI, can substantially increase the sample size by over 157% relative to using only human-annotated data, which results in a dataset larger than the original and improves training and accuracy. This suggests that simpler models, when well-calibrated and properly integrated into such frameworks, can support broader, validity-guaranteed conclusions in educational assessment settings.

The results obtained in this paper have broad implications for large-scale and AI-mediated learning environments, where many learners require assessment and guidance, and human feedback is impractical (Swiecki et al., 2022). In such contexts, a small, well-annotated dataset can be used to make valid predictions on larger unlabeled data, reducing training requirements, improving prediction quality, and enabling large-scale assessments with validity guarantees.

7 Limitations

This study aimed to explore the potential of Prediction-Powered Inference (PPI) to extend a small set of high-quality, human-scored conversational responses using less reliable data generated by large language models (LLMs) and simpler machine learning models. While PPI offers a statistically grounded framework for leveraging such predictions, it assumes that the labeled data are independently and identically distributed (i.i.d.) from a normal distribution. Although our labeled sample size ($n = 445$) may appear limited, each session includes, on average, over eight student turns, providing a richer source of information per data point. Nonetheless, larger samples of gold-standard data should be examined in future work to validate and generalize the findings presented here. Expanding the dataset to include a broader range of learner proficiencies could also provide further insights into the robustness and adaptability of the proposed approach.

8 Ethical considerations

In this study, we caution against the use of LLM-generated outputs as ground-truth data in educa-

tional settings, emphasizing the risks associated with treating such predictions as authoritative. Nevertheless, we acknowledge the potential of LLMs in low-stakes educational scenarios, particularly for generating synthetic data or instructional materials that can support learning.

It is important to note that this work assumes human annotations as the gold standard. However, this assumption should be approached with caution, as human judgments are subject to both cognitive (Gautam and Srinath, 2024) and socio-cultural biases (Huang and Yang, 2023), which are often context-dependent and may impact the reliability of reference scores.

References

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

Anastasios N Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I Jordan, and Tijana Zrnic. 2023. Prediction-powered inference. *Science*, 382(6671):669–674.

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

Serge Bibauw, Thomas François, and Piet Desmet. 2019. Discussing with a computer to practice a foreign language: Research synthesis and conceptual framework of dialogue-based call. *Computer Assisted Language Learning*, 32(8):827–877.

Pierre Boyeau, Anastasios N Angelopoulos, Nir Yosef, Jitendra Malik, and Michael I Jordan. 2024. AutoEval done right: Using synthetic data for model evaluation. *arXiv preprint arXiv:2403.07008*.

Steven Burrows, Iryna Gurevych, and Benno Stein. 2015. The eras and trends of automatic short answer grading. *International journal of artificial intelligence in education*, 25:60–117.

Ivi Chatzi, Eleni Straitouri, Suhas Thejaswi, and Manuel Rodriguez. 2024. Prediction-powered ranking of large language models. *Advances in Neural Information Processing Systems*, 37:113096–113133.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press, Cambridge.

Scott Crossley, Amanda Clevinger, and YouJin Kim. 2014. The role of lexical properties and cohesive devices in text integration and their effect on human ratings of speaking proficiency. *Language Assessment Quarterly*, 11(3):250–270.

Scott A Crossley, Tom Salsbury, Danielle S McNamara, and Scott Jarvis. 2011. Predicting lexical proficiency in language learner texts using computational indices. *Language testing*, 28(4):561–580.

Mary K Enright and Thomas Quinlan. 2010. Complementing human judgment of essays written by english language learners with e-rater® scoring. *Language Testing*, 27(3):317–334.

Naiyi Xie Fincham and Aitor Arronte Alvarez. 2024. Using large language models (LLMs) to facilitate l2 proficiency development through personalized feedback and scaffolding: An empirical study. In *Proceedings of the International CALL Research Conference*, volume 2024, pages 59–64.

Karën Fort. 2016. *Collaborative annotation for reliable natural language processing: Technical and sociological aspects*. John Wiley & Sons.

Sanjana Gautam and Mukund Srinath. 2024. Blind spots and biases: Exploring the role of annotator cognitive biases in NLP. In *Proceedings of the Third Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 82–88, Mexico City, Mexico. Association for Computational Linguistics.

Dominik Glandorf, Peng Cui, Detmar Meurers, and Mrinmaya Sachan. 2025. Grammar control in dialogue response generation for language learning chatbots. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9820–9839, Albuquerque, New Mexico. Association for Computational Linguistics.

Kristina Gligorić, Tijana Zrnic, Cinoo Lee, Emmanuel J Candès, and Dan Jurafsky. 2024. Can unconfident LLM annotations be used for confident conclusions? *arXiv preprint arXiv:2408.15204*.

Akshay Goel, Almog Gueta, Omry Gilon, Chang Liu, Sofia Erell, Lan Huong Nguyen, Xiaohong Hao, Bolous Jaber, Shashir Reddy, Rupesh Kartha, et al.

2023. Llms accelerate annotation for medical information extraction. In *Machine Learning for Health (ML4H)*, pages 82–100. PMLR.

Anisha Gunjal, Jihan Yin, and Erhan Bas. 2024. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18135–18143.

Fei Huang, Arun Ahuja, Doug Downey, Yi Yang, Yuhong Guo, and Alexander Yates. 2014. Learning representations for weakly supervised natural language processing tasks. *Computational Linguistics*, 40(1):85–120.

Jing Huang and Diyi Yang. 2023. Culturally aware natural language inference. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7591–7609, Singapore. Association for Computational Linguistics.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30.

Zixuan Ke and Vincent Ng. 2019. Automated essay scoring: A survey of the state of the art. In *IJCAI*, volume 19, pages 6300–6308.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Diane Litman, Helmer Strik, and Gad S Lim. 2018. Speech technologies and the assessment of second language speaking: Approaches, challenges, and opportunities. *Language Assessment Quarterly*, 15(3):294–309.

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

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

Ellis B Page. 1968. The use of the computer in analyzing student essays. *International review of education*, 14:210–225.

Yudi Pawitan and Chris Holmes. 2025. Confidence in the reasoning of large language models. *Harvard Data Science Review*, 7(1).

Vikram Ramanarayanan, Matthew Mulholland, and Yao Qian. 2019. Scoring interactional aspects of human-machine dialog for language learning and assessment using text features. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 103–109, Stockholm, Sweden. Association for Computational Linguistics.

Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2017. Snorkel: rapid training data creation with weak supervision. *Proc. VLDB Endow.*, 11(3):269–282.

Wendi Ren, Yinghao Li, Hanting Su, David Kartchner, Cassie Mitchell, and Chao Zhang. 2020. Denoising multi-source weak supervision for neural text classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3739–3754, Online. Association for Computational Linguistics.

Pedro Uria Rodriguez, Amir Jafari, and Christopher M Ormerod. 2019. Language models and automated essay scoring. *arXiv preprint arXiv:1909.09482*.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2024. ARES: An automated evaluation framework for retrieval-augmented generation systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 338–354, Mexico City, Mexico. Association for Computational Linguistics.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2023. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *The Twelfth International Conference on Learning Representations*.

Mark D Shermis and Jill Burstein. 2013. *Handbook of automated essay evaluation*. Routledge, New York.

Mark D Shermis and Jill C Burstein. 2003. *Automated Essay Scoring: A Cross-disciplinary Perspective*. Routledge, New York.

Ravid Shwartz-Ziv and Amitai Armon. 2022. Tabular data: Deep learning is not all you need. *Information Fusion*, 81:84–90.

Ryan Smith, Jason A Fries, Braden Hancock, and Stephen H Bach. 2024. Language models in the loop: Incorporating prompting into weak supervision. *ACM/JMS Journal of Data Science*, 1(2):1–30.

Anders Søgaard. 2022. *Semi-supervised learning and domain adaptation in natural language processing*. Springer Nature.

Zachari Swiecki, Hassan Khosravi, Guanliang Chen, Roberto Martinez-Maldonado, Jason M Lodge, Sandra Milligan, Neil Selwyn, and Dragan Gašević. 2022. Assessment in the age of artificial intelligence. *Computers and Education: Artificial Intelligence*, 3:100075.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024a. Large language models for data annotation: A survey. *arXiv preprint arXiv:2402.13446*.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024b. Large Language Models for data annotation and synthesis: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 930–957, Miami, Florida, USA. Association for Computational Linguistics.

Tamara P Tate, Jacob Steiss, Drew Bailey, Steve Graham, Youngsun Moon, Daniel Ritchie, Waverly Tseng, and Mark Warschauer. 2024. Can AI provide useful holistic essay scoring? *Computers and Education: Artificial Intelligence*, 7:100255.

Georgelle Thomas, R Derald Hartley, and J Peter Kincaid. 1975. Test-retest and inter-analyst reliability of the automated readability index, flesch reading ease score, and the fog count. *Journal of Reading Behavior*, 7(2):149–154.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, Singapore. Association for Computational Linguistics.

Gladys Tyen, Mark Brenchley, Andrew Caines, and Paula Buttery. 2022. Towards an open-domain chatbot for language practice. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 234–249, Seattle, Washington. Association for Computational Linguistics.

Xiaoming Xi. 2010. Automated scoring and feedback systems: Where are we and where are we heading? *Language Testing*, 27(3):291–300.

Kevin P. Yancey, Geoffrey Laflair, Anthony Verardi, and Jill Burstein. 2023. Rating short L2 essays on the CEFR scale with GPT-4. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 576–584, Toronto, Canada. Association for Computational Linguistics.

Steven Kester Yuwono, Biao Wu, and Luis Fernando D'Haro. 2019. Automated scoring of chatbot responses in conversational dialogue. In *9th International Workshop on Spoken Dialogue System Technology*, pages 357–369. Springer.

Jieyu Zhang, Yue Yu, Yinghao Li, Yujing Wang, Yaming Yang, Mao Yang, and Alexander Ratner. 2021. WRENCH: A comprehensive benchmark for weak supervision. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Dawei Zhu, Xiaoyu Shen, Marius Mosbach, Andreas Stephan, and Dietrich Klakow. 2023. Weaker than you think: A critical look at weakly supervised learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14229–14253, Toronto, Canada. Association for Computational Linguistics.

A Rubric

A.1 Vocabulary Range (B1-B2)

Score	Description of the proficiency level
1	Has a good range of vocabulary related to familiar topics and everyday situations. Has sufficient vocabulary to express themselves with some circumlocutions on most topics pertinent to their everyday life such as family, hobbies and interests, work, travel and current events.
2	Has a good range of vocabulary for matters connected to their field and most general topics. Can vary formulation to avoid frequent repetition, but lexical gaps can still cause hesitation and circumlocution. Can produce appropriate collocations of many words/signs in most contexts fairly systematically. Can understand and use much of the specialist vocabulary of their field but has problems with specialist terminology outside it.
3	Can understand and use the main technical terminology of their field, when discussing their area of specialisation with other specialists.

Table 2: Vocabulary Range Rubric (B1-B2)

A.2 Grammatical Accuracy (B1-B2)

Score	Description of the proficiency level
1	Uses reasonably accurately a repertoire of frequently used routines and patterns associated with more predictable situations.
2	Communicates with reasonable accuracy in familiar contexts; generally good control, though with noticeable mother-tongue influence. Errors occur, but it is clear what they are trying to express.
3	Has a good command of simple language structures and some complex grammatical forms, although they tend to use complex structures rigidly with some inaccuracy.
4	Good grammatical control; occasional slips or non-systematic errors and minor flaws in sentence structure may still occur, but they are rare and can often be corrected in retrospect.

Table 3: Grammatical Accuracy Rubric (B1-B2)

B Accuracy for gold standard set

Model	Proficiency level	Accuracy
XGBoost	Vocabulary Range	74.1
XGBoost	Grammatical accuracy	71.3
LightGBM	Vocabulary Range	73.6
LightGBM	Grammatical accuracy	71
GPT-4o	Vocabulary Range	62.5
GPT-4o	Grammatical accuracy	61.2
GPT-4o-mini	Vocabulary Range	60.3
GPT-4o-mini	Grammatical accuracy	58.9

Table 4: Accuracy values for each of the models used tested on the gold standard set only.

C Prompts

Score the following text from a conversation of an intermediate English language student (B1-B2 on CEFR).
Provide the score as an integer and the probability as a float associated with the options in the 'ScoringTexts' function.
Text: text

```
class ScoringTexts(BaseModel):
  #CEFR vocabulary range.
  vocabulary_range: int = Field(description="Select the option that best describes
      the text."
                                          "Option 1. Has a good range of vocabulary
                                              related to familiar topics and
                                             everyday situations."
                                           "Has sufficient vocabulary to express
                                              themselves with some circumlocutions
                                               on most topics "
                                            "pertinent to their everyday life such
                                                as family, hobbies and interests,
                                                work, travel and current events."
                                           "Option 2. Has a good range of
                                              vocabulary for matters connected to
                                              their field and most general topics.
                                              "
                                           "Can vary formulation to avoid frequent
                                              repetition, but lexical gaps can
                                              still cause hesitation"
                                           " and circumlocution."
                                           "Can produce appropriate collocations of
                                               many words/signs in most contexts
                                              fairly systematically."
                                           "Can understand and use much of the
                                              specialist vocabulary of their field
                                               but has problems with "
                                           "specialist terminology outside it."
                                           "Option 3. Can understand and use
                                              technical terminology when
                                              discussing "
                                           "areas of specialization. Have access to
                                               specialized vocabulary in relation
                                              to the topic.")

  vocabulary_range_proba: float = Field(description="Express in the form of a
      probability the confidence on the vocabulary range score given.")


  #measures of grammatical accuracy as per CEFR
  grammatical_accuracy: int = Field(description="Select the option that best
      describes the text."
                                             "Option 1. Uses reasonably accurately
                                                a repertoire of frequently used
                                                routines and patterns "
                                             "associated with more predictable
                                                situations. "
                                             "Option 2. Communicates with
                                                reasonable accuracy in familiar
                                                contexts; generally good control,
                                                "
                                              "though with noticeable mother-
                                                 tongue influence."
                                             "Errors occur, but it is clear what
                                                they are trying to express."
                                             "Option 3. Has a good command of
                                                simple language structures and
                                                some complex grammatical forms, "
                                              "although they tend to use complex
```

```
                                                    structures rigidly with some
                                                    inaccuracy."
                                                "option 4. Good grammatical control;
                                                    occasional slips or non-systematic
                                                    errors and minor flaws "
                                                "in sentence structure may still
                                                    occur, "
                                                "but they are rare and can often be
                                                    corrected in retrospect.")

    grammatical_accuracy_proba: float = Field(description="Express in the form of a
        probability the confidence on the grammatical accuracy score given.")
```