Beyond the Score: Uncertainty-Calibrated LLMs for Automated Essay Assessment

Anonymous EMNLP submission

Abstract

Automated Essay Scoring (AES) systems now attain near–human agreement on public benchmarks, yet real-world adoption—especially in high-stakes examinations—remains limited. A principal obstacle is that most models output a *single* score without any accompanying measure of confidence or explanation. We address this gap with *conformal prediction*, a distribution-free wrapper that equips any classifier with set-valued outputs enjoying formal coverage guarantees. Two open-weight large language models—**Llama-3 8B** and **Qwen-2.5 3B**—are fine-tuned on three diverse corpora (ASAP, TOEFL11, Cambridge-FCE) and calibrated at a 90% risk level. Reliability is assessed with **UAcc**, an uncertainty-aware accuracy that rewards models for being both correct and concise. To our knowledge, this is the first work to combine conformal prediction and UAcc for essay scoring. The calibrated models consistently meet the coverage target while keeping prediction sets compact, demonstrating that trustworthy, uncertainty-aware AES is already feasible with mid-sized, open source LLMs and paving the way for safer human-in-the-loop marking.

1 Introduction

Automated Essay Scoring (AES) has evolved rapidly—from linear regressors built on handcrafted features (Phandi et al., 2015), through CNN–LSTM hybrids that capture local and long-range coherence (Taghipour and Ng, 2016), to transformer encoders such as R2BERT that pair BERT representations with joint regression–ranking objectives and reach state-of-the-art agreement on ASAP essays (Yang et al., 2020). The latest step is the move to open-weight large language models (LLMs): lightly tuned Llama variants now approach human-human agreement on several AES benchmarks (Xiao et al., 2025).

Headline accuracy, however, is not enough for high-stakes settings such as TOEFL or the Cambridge First Certificate, where a single mis-scored script can determine admission or visa status. Exam boards need calibrated confidence. Common approaches to measuring uncertainty include Monte-Carlo dropout (Gal and Ghahramani, 2016), deep ensembles (Lakshminarayanan et al., 2017) and Bayesian neural networks(Goan and Fookes, 2020). These methods are effective but either multiply inference cost or offer no finite-sample guarantees.

Conformal prediction (CP) (Angelopoulos and Bates, 2021) provides such guarantees by wrapping any classifier with a *set-valued* output that contains the true label with user-chosen probability. CP has improved reliability in tasks from surrogate models (Gopakumar et al., 2024) to question answering, yet it has not been applied to AES, and no study has linked calibration quality to scoring usefulness. We bridge that gap with **UAcc**—an uncertainty-aware accuracy that rewards models that are correct and selective (Ye et al., 2024).

In this paper, we fine-tune two state-of-the-art LLMs—**Llama-3 8B** and **Qwen-2.5 3B**—on three public essay corpora (ASAP (Kaggle, 2012), TOEFL11 (Daniel Blanchard, 2014), Cambridge-FCE (Yannakoudakis et al., 2011)). Each scorer is then calibrated with conformal prediction so that its prediction set is guaranteed, by construction, to contain the true score in at least 90 % of future essays. We evaluated these calibrated models with the uncertainty-aware accuracy **UAcc**, alongside standard accuracy and quadratic-weighted κ. Across all corpora, the models meet the 90% coverage guarantee while keeping prediction sets tight, showing that uncertainty-aware AES is already practical with mid-sized, openly licensed LLMs. By uniting modern language models, distribution-free calibration, and an uncertainty-sensitive metric, we provide the first comprehensive picture of trustworthy essay scoring across multiple proficiency tests.

2 Background

This work combines a standard essay-scoring model with *conformal prediction* so that every prediction comes with a statistically valid notion of confidence.

2.1 Essay-scoring task

An essay x must receive one label y from a fixed set of K possible scores (e.g. the integers 2–12 for ASAP, or the three bands *low/med/high* for TOEFL). A neural scorer f takes the essay text and outputs a probability for each label; we denote that distribution by $\hat{p}(y \mid x)$.

2.2 Conformal prediction (CP)

Conformal prediction turns those probabilities into a **prediction set** $C_\alpha(x) \subseteq \{1, \ldots, K\}$ that is guaranteed to contain the true score with high probability. Formally, for a user-chosen risk level α (we use $\alpha = 0.1$), CP ensures

$$\Pr[y \in C_\alpha(x)] \;\geq\; 1 - \alpha \tag{1}$$

so the true score falls outside the set in at most 10% of future essays.

How conformal sets are constructed. To build a prediction set using conformal prediction, the data is first split into three parts: a training set for fitting the model, a calibration set for estimating uncertainty, and a test set for evaluation.

For a given model f and input essay x, the score assigned to each possible label y is defined as

$$s(x, y) = 1 - \hat{p}(y \mid x) \tag{2}$$

where $\hat{p}(y \mid x)$ is the model's predicted probability. This is known as the least-ambiguous classifier (LAC) score—lower scores indicate higher confidence.

Using the calibration set, the conformal algorithm computes a threshold q_α such that at most an α fraction of calibration scores exceed it. Then, for any new essay x, the prediction set is formed by including all labels with scores below this threshold:

$$C_\alpha(x) = \{y \in \mathcal{Y} \mid s(x, y) \leq q_\alpha\} \tag{3}$$

This guarantees that the prediction set contains the true label with probability at least $1 - \alpha$.

2.3 Metrics

We evaluate models using both standard and uncertainty-aware criteria. In addition to accuracy and quadratic-weighted kappa (QWK), we report three key metrics specific to conformal prediction: (i) **Coverage**, the proportion of test essays for which the true label is contained in the prediction set $C_\alpha(x)$; (ii) **Average set size**, measuring how many labels are typically included—smaller is better; and (iii) **UAcc** which balances correctness and conciseness via

$$UAcc = Accuracy \times \sqrt{\frac{K}{avg.\,|C_\alpha(x)|}} \tag{4}$$

where K is the number of classes. UAcc penalises large or overly cautious prediction sets, rewarding models that are both accurate and selective.

3 Experimental Setup

3.1 Models and tokenisation

We experiment with two openly licensed generative LLMs: **Llama-3 8B** (Dubey et al., 2024) and **Qwen-2.5 3B** (Yang et al., 2024). Both are loaded via HuggingFace Transformers with 4-bit quantisation; special tokens and maximum context length follow the model cards. For each corpus we prepend a short instruction—*"Read the essay and output a single score:"*—and rely on the tokenizer to convert either the integer label (ASAP) or the band token *low/medium/high* (TOEFL11, FCE) into a single ID, so that the final token distribution can be treated as a 3- or 11-way classifier without adding new parameters.

3.2 Fine-tuning

Training is performed on a single Nvidia A100-40GB GPU for eight epochs with AdamW (learning rate 1×10^{-5}) We use a global batch size of 8 and truncate inputs at 256 tokens. A fixed random seed (42) is used ensure reproducibility.

3.3 Calibration and test split

After fine-tuning, the original validation + test portion of each corpus is split once into equal-sized *calibration* and test sets (15 % / 15 % of the full data; exact counts in Table 1). Calibration essays never influence model weights.

3.4 Conformal prediction

For every essay–label pair we compute the least-ambiguous classifier score $s(x, y) = 1 - \hat{p}(y \mid x)$,

where $\hat{p}$ is the model's softmax probability. The $(1-\alpha)$ quantile of these scores on the calibration set with $\alpha = 0.1$ yields the threshold q_α. At inference time we return all labels whose scores fall below q_α, guaranteeing that the prediction set contains the true label in at least 90 % of future essays.

3.5 Evaluation metrics

We report conventional accuracy and quadratic-weighted κ (QWK) together with three uncertainty-aware measures introduced in Section 2: **Coverage**, the empirical proportion of essays whose true label lies in the prediction set; **Average set size**, a proxy for informativeness; and **UAcc**, which trades off accuracy against set size.

3.6 Datasets

Table 1 lists the three corpora and the statistics derived from our 70 / 15 / 15 train–calibration–test split.

ASAP Prompt 1 Essays written by secondary-school students and graded on an eleven-point scale (2–12). We keep the full scale for as retaining an intermediate label space allows us to study how uncertainty behaves when the number of possible scores increases—something neither TOEFL11 (3 classes) nor FCE (3 bands) can reveal.

TOEFL11 Internet-based TOEFL essays pre-labelled *low*, *medium* or *high*.

Cambridge-FCE Scripts scored holistically 1–40. To align with TOEFL11 and keep prediction sets interpretable, we divide the range into equal-width thirds—1–18 (*low*), 19–30 (*medium*) and 31–40 (*high*). This heuristic balances the three classes and prevents prediction sets from ballooning to forty labels; exploring finer buckets is left for future work.

4 Results and Discussion

Across all three corpora, the calibrated Llama models achieve the highest quadratic-weighted κ, confirming that stronger back-bones still translate into better agreement with human graders even after quantisation and LoRA fine-tuning. Crucially, they do so while returning the *tightest* prediction sets: roughly 2.7 labels on the 11-way ASAP rubric and fewer than two labels on the three-class TOEFL11 and FCE tasks. Those concise sets lift UAcc above competing systems that share the same point accuracy. In other words, the Llama scorers are not merely correct; they are confident enough to commit to a smaller subset of possible scores, which reduces the burden on any downstream human reviewer.

Because UAcc rescales accuracy by $\sqrt{K/|C|}$, a system can gain either by raising raw accuracy or by shrinking its prediction sets. On ASAP, Llama-2 and Qwen differ by only four accuracy points (0.54 vs 0.50), yet Llama's sets are 0.8 labels tighter, boosting UAcc from 0.88 to 1.08. In practice that means nearly 20 % fewer essays would be flagged for manual review at the same error rate—an operationally significant saving.

Accuracy can be misleading whenever the label distribution is skewed. F1 gives equal weight to each class, revealing whether a model simply exploits the majority label or performs consistently across bands. In our results the gap between accuracy and F1 is small, confirming that the calibrated LLMs do not over-predict a single band; nonetheless, reporting F1 guards against potential imbalance and strengthens the claim that the models generalise across proficiency levels.

Empirical coverage lies within one percentage point of the 90% target on every dataset, demonstrating that a single conformal wrapper generalises from an 11-point rubric (ASAP) to 3-labeled (TOEFL11, FCE) despite the shift in prompt style, score range and proficiency level. The larger prediction sets observed on ASAP reflect the richer label space: with eleven possible scores the model must sometimes hedge between adjacent grades, a phenomenon less common in the three-band corpora.

The absolute QWK numbers on FCE are markedly lower than on TOEFL11, even though both datasets use the same low/medium/high mapping. Two factors help to explain the gap. First, the FCE essays are mapped *post-hoc* from a 40-point holistic scale, and quadratic penalises any band disagreement proportionally to the original distance on that underlying scale; a one-band slip therefore receives a much heavier penalty than in TOEFL11, whose native rubric already has three discrete labels. Second, the FCE corpus is almost one order of magnitude smaller than TOEFL11, magnifying the impact of label noise and leaving less data for both fine-tuning and calibration. Taken together, these artefacts depress QWK even when coverage and

Corpus	Train	Cal	Test	Labels
ASAP P1	1 248	268	267	11-way (2–12)
TOEFL11	8 470	1 815	1 815	low / med / high
Cambridge-FCE	1 742	373	373	low / med / high

Table 1: Dataset sizes after a 70 / 15 / 15 train–calibration–test split. These are the number of essays in each split

Dataset	Model	QWK	Acc.	F1	Coverage	Avg. $\lvert C \rvert$	UAcc
ASAP P1	Qwen-2.5 3B	0.69	0.50	0.45	0.91	3.51	0.88
	Llama-2 7B	**0.82**	**0.54**	**0.52**	**0.91**	**2.74**	**1.08**
	Llama-3 8B	0.80	**0.54**	0.51	**0.91**	2.81	1.07
TOEFL11	Qwen-2.5 3B	0.69	0.77	0.76	0.89	1.32	1.16
	Llama-3 8B	**0.70**	0.77	**0.77**	0.89	**1.29**	**1.17**
Cambridge-FCE	Qwen-2.5 3B	0.16	0.65	0.62	**0.95**	2.30	0.74
	Llama-3 8B	**0.28**	**0.66**	**0.64**	0.88	**1.74**	**0.87**

Model (paper)	Training epochs / runs	QWK (ASAP Prompt 1)
EASE (SVR) (Taghipour and Ng, 2016)	—	0.781
LSTM (10×) (Taghipour and Ng, 2016)	10 runs	0.808
Ensemble CNN+LSTM (Taghipour and Ng, 2016)	10 runs	0.821
R2BERT (Yang et al., 2020)	30 epochs	0.817
Fine-tuned GPT-3.5 (Xiao et al., 2025)	10 epochs	0.740
Fine-tuned LLaMA-3 (2-pt) (Xiao et al., 2025)	10 epochs	0.714
Our LLaMA-3 8B (8 ep)	8 epochs	**0.800**
Our LLaMA-2 7B (8 ep)	8 epochs	**0.823**

Table 2: Top: calibrated performance on three corpora. Bottom: published ASAP Prompt 1 baselines vs. our systems.

UAcc remain competitive.

Overall, these findings show that mid-sized, openly licensed LLMs already deliver high scoring accuracy together with calibrated, interpretable uncertainty which are key prerequisites for deployment in high-stakes assessment. The consistent edge of Llama-3 over its smaller Qwen counterpart confirms that parameter count and pre-training data still matter, yet the margin is small enough to keep lower-footprint models in serious contention wherever hardware or licensing constraints apply.

5 Conclusion

We set out to answer whether modern large language models can score essays *and* express calibrated uncertainty in a way that is practical for high-stakes assessment. By wrapping two LLMs—Llama-3 8B and Qwen-2.5 3B—with conformal prediction and judging them with the uncertainty-aware metric **UAcc**, we showed that a single, distribution-free calibration step delivers near-perfect coverage (90 %) across three very different corpora. The stronger Llama backbone achieves the best trade-off between agreement with human graders (QWK) and prediction-set tightness, yet the gap to the smaller Qwen model is modest—evidence that trustworthy AES does not require flagship-scale models. Taken together, these results provide the first end-to-end demonstration that mid-sized, openly licensed LLMs can power calibrated, human-in-the-loop essay scoring systems today, while laying the groundwork for future studies on model size, finer FCE banding, and rubric-aware prompting.

Future work will probe the trade-off of model size and performance more systematically: we plan to train a spectrum of model sizes (1B to 13B) from several families to quantify when accuracy and UAcc begin to show diminishing returns. On the data side, we will experiment with finer-grained buckets for the FCE corpus and, more generally, with ordinal-aware conformal scores that respect the underlying scale. Finally, we intend to condition prompts on essay characteristics—length, discourse structure to see whether rubric-aware prompting can tighten prediction sets still further without sacrificing coverage.

Limitations

This study is confined to English and to three public essay corpora, two of which we deliberately reduce to a three-band rubric for comparability. Although conformal prediction delivers the promised 90 % coverage under these conditions, the guarantee relies on calibration and test data being exchangeable; topic drift, candidate demographics or language transfer effects in a real exam session could weaken reliability. Our choice of equal-width bands for the 1–40 Cambridge-FCE scale is a heuristic that balances class counts but may mask finer proficiency distinctions. Likewise, we retain ASAP's full 11-point rubric to explore class-rich uncertainty, yet that decision limits direct comparison across datasets.

From a practical standpoint, even 4-bit LoRA fine-tuning of an 8B-parameter model requires a high-end GPU; institutions with modest hardware may still prefer smaller models. Finally, while calibrated prediction sets indicate *how sure* the model is, they do not explain *why* a script is low, medium or high; integrating rubric-aligned rationales is an important next step toward truly interpretable AES.

References

Anastasios N. Angelopoulos and Stephen Bates. 2021. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *CoRR*, abs/2107.07511.

Derrick Higgins Aoife Cahill Martin Chodorow Daniel Blanchard, Joel Tetreault. 2014. Ets corpus of non-native written english (toefl11). Linguistic Data Consortium, Philadelphia. LDC Catalog No. LDC2014T06.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, New York, New York, USA. PMLR.

Ethan Goan and Clinton Fookes. 2020. Bayesian neural networks: An introduction and survey.

Vignesh Gopakumar, Ander Gray, Joel Oskarsson, Lorenzo Zanisi, Stanislas Pamela, Daniel Giles, Matt Kusner, and Marc Deisenroth. 2024. Uncertainty quantification of pre-trained and fine-tuned surrogate models using conformal prediction.

Kaggle. 2012. Asap automated essay scoring data set. `https://www.kaggle.com/c/asap-aes`.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6405–6416, Red Hook, NY, USA. Curran Associates Inc.

Peter Phandi, Kian Ming A. Chai, and Hwee Tou Ng. 2015. Flexible domain adaptation for automated essay scoring using correlated linear regression. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 431–439, Lisbon, Portugal. Association for Computational Linguistics.

Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1882–1891, Austin, Texas. Association for Computational Linguistics.

Changrong Xiao, Wenxing Ma, Qingping Song, Sean Xin Xu, Kunpeng Zhang, Yufang Wang, and Qi Fu. 2025. Human-ai collaborative essay scoring: A dual-process framework with llms. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, LAK '25, page 293–305, New York, NY, USA. Association for Computing Machinery.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and Zhihao Fan. 2024. Qwen2 technical report.

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: EMNLP 2020*, pages 1560–1569, Online. Association for Computational Linguistics.

Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. 2011. A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 180–189, Portland, Oregon, USA. Association for Computational Linguistics.

Fanghua Ye, Yang MingMing, Jianhui Pang, Longyue Wang, Derek F Wong, Yilmaz Emine, Shuming Shi, and Zhaopeng Tu. 2024. Benchmarking llms via uncertainty quantification. *arXiv preprint arXiv:2401.12794*.