Bounding Boxes to Improve Small Language Model Performance on Vision-Based Grading Tasks

Lachlan McGinness[1]

[1]Australian National University, Acton, ACT, Australia

Abstract

The deployment of Small Language Models (SLMs) in educational settings offers significant advantages in terms of privacy, cost, and scalability. However, SLMs often struggle with complex vision-based tasks, such as grading handwritten student exams, due to the high computational cost of processing large images and the visual distractions present on a full page. In this paper, we investigate whether cropping student responses using bounding boxes can improve the accuracy and computational efficiency of SLMs on a short-answer grading task. Using a dataset of scanned handwritten responses from the 2025 Australian Physics Olympiad, we evaluate the performance of several models ranging from 4B to 72B parameters under varying conditions of Chain of Thought (CoT) prompting and image cropping. Our results demonstrate that using bounding boxes significantly improves grading accuracy and reduces computational cost (FLOPs) across models. We conclude that bounding boxes are a crucial pre-processing step for deploying SLMs in large-scale, vision-based educational assessments.

Keywords

Small Language Models, Automated Grading, Vision-Language Models, Bounding Boxes, AI for Education

1. Introduction

The integration of Artificial Intelligence in Education (AIED) has rapidly expanded the use of Large Language Models (LLMs) including Vision Language Models (VLMs) for tasks such as automated short-answer grading [1, 2, 3, 4, 5, 6, 7, 8, 9]. These studies use large, cloud-hosted, proprietary models. In parallel, a growing body of work has begun to examine Small Language Models (SLMs)[1] as alternatives that can be deployed locally on consumer-grade hardware [10, 11, 12, 13]. These studies indicate that SLMs can achieve competitive performance on a range of educational generation and evaluation tasks while offering important advantages in privacy, cost, and institutional control over data. These properties are particularly attractive in schools, where sensitive student data, regulatory constraints, and limited budgets could prevent the use of proprietary cloud-based LLMs.

A particularly demanding application of language models in education is the grading of *handwritten* student work. In contrast to textual short-answer grading, where the student's response is already digitised, handwritten exams require the model to locate the answer on the page, interpret crossed-out text, and ignore unrelated content from other questions.

VLMs tokenise images by partitioning them into patches that are projected into the language model's embedding space, so that the number of *visual* tokens scales (approximately) linearly with the image area. Modern dynamic-resolution architectures such as Gemma3 and Qwen2.5 preserve native aspect ratios and produce visual token counts that are roughly proportional to the number of 14×14 pixel patches in the input [14, 15]. A full A3 double-page exam scan can therefore consume thousands of visual tokens, and recent analyses have shown that vision-encoder cost grows rapidly with input resolution, often dominating the computational cost of a forward pass [16]. Under the standard $C \approx 2nN$ approximation for inference compute [17, 18], where n is the number of tokens processed and N is the number of active parameters, this directly translates into a higher number of FLOPs (computational cost) per graded response.

VLMs can be easily distracted by content-rich pages which contain a large number of visual features

SLM4ED'26: The 1st Workshop of Small Language Models for Education (SLM4ED), June 28, 2026, Seoul, Republic of Korea
lachlan.mcginness@anu.edu.au (L. McGinness)
© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

[1]typically taken to mean open-weights models with up to about 3×10^{10} parameters

CEUR-WS.org/Vol-4296/short1.pdf

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-10-19

that are irrelevant to any one sub-question; for example other questions or scribbles. Therefore SLMs may benefit significantly from reducing the visual search problem before the model sees the image. Classical document layout analysis has long relied on cropping a region of interest as a preprocessing step prior to OCR and we hypothesise that selectively decoding only regions relevant to a query will improve both accuracy and efficiency.

In this short study, we test this intuition empirically in an educational setting. Using 126 scanned handwritten responses to Question 1a of the 2025 Australian Physics Olympiad, we evaluate eight open-source VLMs ranging from 4B to 72B parameters with two experimental variables: (i) using of CoT prompting and (ii) whether the model is given the full A3 double-page spread or only a bounding-box crop of the relevant question/answer region. We measure both grading accuracy against human markers and the total inference FLOPs implied by the input/output token counts.

2. Methodology

For our experiments we used Question 1a of the 2025 Australian Physics Olympiad exam. Students completed this exam on paper supervised by their teachers. As students registered online, they were provided with a participant information sheet and able to tick a box to opt-in to consenting for their responses to be used for the research. In total, 588 of the 1357 students who sat the exam consented, and their exams were scanned as double-page spreads (landscape A3 pages). To reduce the computational cost and time to run this preliminary experiment, 126 of the 588 student exam papers were chosen randomly as the dataset for each of our four experimental conditions. Question 1a was a short answer question, worded as follows:

> Imagine a human walking on perfectly flat ground. With each step they accelerate and decelerate. Which force is responsible for this acceleration and deceleration?

The answer is simply ‘friction’ or ‘friction force’. The main challenge for grading this question is reading the student’s handwritten response, as very little reasoning is required.

2.1. Experimental Conditions

In this experiment, we investigated two sets of independent variables: Chain of Thought (CoT) prompting and bounding boxes.

Chain of Thought Prompting To investigate the influence of chain of thought prompting, each of the eight models were used to grade the 126 responses with two different prompts. The prompt for the CoT condition was [2]:

> “This answer page contains a student response to a question about snails. Please read the student’s answer for Question 1a near the top of the page. If they wrote ‘friction force’, ‘friction’, or something similar please output 1. Otherwise output 0. If the student wrote friction and then crossed it out and then wrote something else write 0. Think this through step by step but then ensure ‘0’ or ‘1’ is the last thing that you write.”

The prompt for the no-CoT condition was:

> This answer page contains a student response to a question about snails. Please read the student’s answer for Question 1a near the top of the page. If they wrote ‘friction force’, ‘friction’, or something similar please output 1. Otherwise output 0. If the student wrote friction and then crossed it out and then wrote something else write 0. Please don’t write anything else except the 1 or 0 based on the student response. If you do write something else, ensure ‘0’ or ‘1’ is the last thing that you write.

[2]Note that although Question 1a is about human walking, the rest of Question 1 is about snail locomotion and there is a large diagram of a snail immediately below the question.

The distinguishing factor between these prompts is that in CoT, the model is asked to 'think this through step by step'. CoT was one of the most successful early prompt engineering techniques [19]. More recently, Language Models have been trained to natively engage in step by step thinking, so CoT prompts have a lesser [20] or negative effect [21].

Bounding Boxes By inspection of the scanned exams, the researchers determined that pixels corresponding to Question 1a were always in a very similar location. This allowed the researchers to automatically apply bounding boxes to crop and capture the students' answers to the question. In the bounding box condition, only these areas were given to the model to determine the grade. This means that there should be fewer distractions for the model when grading the question. The bounding boxes were quite generous so that if the students wrote off to the side, their response would still successfully be captured, allowing for small variations in paper position during scanning.

In the non-bounding box condition, the entire double-page spread was provided to the model. The model was then required to find the student response to question 1a and extract it. Figure 1 gives an example of the same student response with and without bounding boxes.

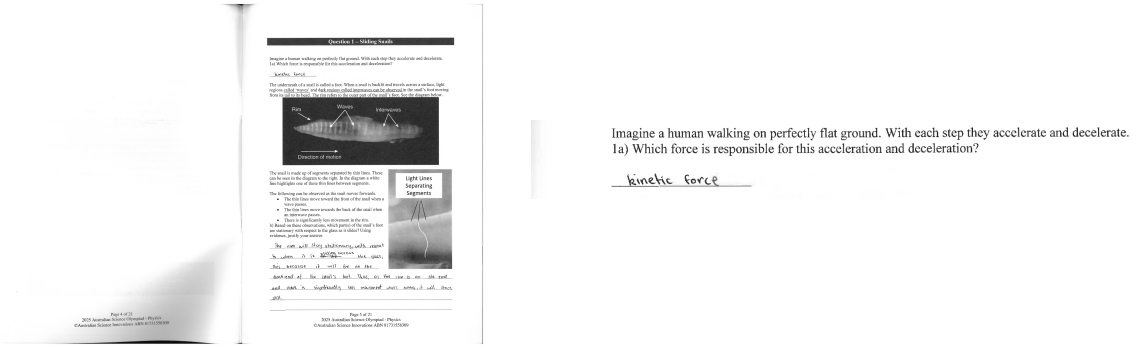

Figure 1: Visual comparison of the bounding box and non-bounding box experimental conditions. **Left:** Original scan of student exam paper. **Right:** Cropped bounding box of the same student exam paper. The bounding box reduces the number of visual features that could confuse smaller models.

Intuitively, the size of an image is approximately proportional to the number of tokens that it takes to encode it. Therefore presenting a model with a smaller picture will require fewer input tokens and reduce the overall computational cost.

In total, there were four experimental conditions: CoT with bounding boxes, CoT without bounding boxes, no-CoT with bounding boxes, and no-CoT without bounding boxes. For each experimental condition, we recorded the raw model response, the number of input tokens, and the number of output tokens. The number of tokens was used to calculate the total FLOPs using the $2nN$ approximation. Since the model's response is binary (0 or 1), accuracy was determined by calculating the fraction of responses for which the model agrees with the human marking team (which had no discrepancies on this question).

2.2. Models Evaluated

The models used in this experiment were open-source and small enough to run on upper-end consumer-grade hardware, making them ideal candidates for local educational deployments. The models evaluated are summarised in Table 1. Note that we included larger models (32B-72B active parameters) as performance ceilings for comparison.

3. Results

Table 2 shows the performance of each of the models in each of the four experimental conditions. Experimental conditions using bounding boxes were the most accurate for all models. With the

Table 1
Vision-language models evaluated in the bounding box and CoT experiments. For Mixture-of-Experts models, active parameters refer to the number engaged per forward pass.

Model	Developer	Active Params	Total Params	Release Date
Llama 4 Scout	Meta	17B	109B	April 2025
Mistral Small 3.2	Mistral AI	24B	24B	June 2025
Gemma 3 27B	Google	27B	27B	March 2025
Gemma 3 12B	Google	12B	12B	March 2025
Gemma 3 4B	Google	4B	4B	March 2025
Qwen2.5-VL 72B	Alibaba	72B	72B	January 2025
Qwen2.5-VL 32B	Alibaba	32B	32B	March 2025
Qwen2.5-VL 7B	Alibaba	7B	7B	January 2025

exception of Mistral 3.2, models did not benefit significantly from CoT prompting. This anomaly may have occurred because Mistral's instruction-following training was significantly different to that of the other models.

Table 2
Accuracy for each model across the four experimental conditions. Values reported as mean ± standard deviation. For most models, the combination of no-CoT prompt with bounding boxes produced the most accurate results.

Condition	Gemma 4B	Gemma 12B	Gemma 27B	Llama4	Mistral3.2	Qwen 7B	Qwen 32B	Qwen 72B
CoT, No BB	0.49 ± 0.04	0.51 ± 0.04	0.80 ± 0.04	0.77 ± 0.04	0.80 ± 0.04	0.65 ± 0.04	0.78 ± 0.04	0.87 ± 0.03
No-CoT, No BB	0.47 ± 0.04	0.65 ± 0.04	0.68 ± 0.04	0.66 ± 0.04	0.51 ± 0.04	0.92 ± 0.02	0.83 ± 0.03	**0.94 ± 0.02**
CoT + BB	0.55 ± 0.04	0.62 ± 0.04	0.85 ± 0.03	0.79 ± 0.04	**0.93 ± 0.02**	0.79 ± 0.04	0.90 ± 0.03	**0.94 ± 0.02**
No-CoT + BB	**0.56 ± 0.04**	**0.74 ± 0.04**	**0.87 ± 0.03**	**0.90 ± 0.03**	0.64 ± 0.04	**0.95 ± 0.02**	**0.94 ± 0.02**	**0.94 ± 0.02**

Figure 2 illustrates model performance as a function of computational cost (in FLOPs). The data is clearly very noisy as performance varies wildly between models, but there are three trends that emerge. Firstly, as expected, increased computational expense increases accuracy across all experimental conditions. Second, there is little difference between the CoT and no-CoT conditions. This is displayed in Figure 2 as CoT + Bounding Boxes (blue) is similar to No-CoT + Bounding Boxes (green). The same pattern applies for the two non-bounding box conditions (purple and red). Finally, there is a noticeable improvement of the bounding box conditions (green and blue) over the non-bounding box conditions (red and purple).

The significance of this difference in accuracy across all models can be determined by taking the average and standard error in the mean across all trials as shown in Table 3.

Table 3
Average performance metrics across all models under different experimental conditions. The uncertainty is the standard error in the mean across all eight models. The reported tokens are the sum of input and output tokens.

Experimental Condition	Accuracy	Average Tokens	Average FLOPs (Trillions)
CoT, Bounding Box	0.817 ± 0.019	694	18
CoT, No Bounding Box	0.734 ± 0.019	1549	45
No CoT, Bounding Box	0.840 ± 0.020	679	17
No CoT, No Bounding Box	0.728 ± 0.024	1463	41

The accuracy is significantly better for cases where bounding boxes were utilised. The difference between the lowest bounding box accuracy and highest non-bounding box accuracy is 3.08 times the squared quadrature of the standard errors, corresponding to significance at the 99.8% confidence level. Furthermore, bounding boxes reduced the number of tokens required by a significant margin compared to parsing the full image, resulting in substantially less computational expense.

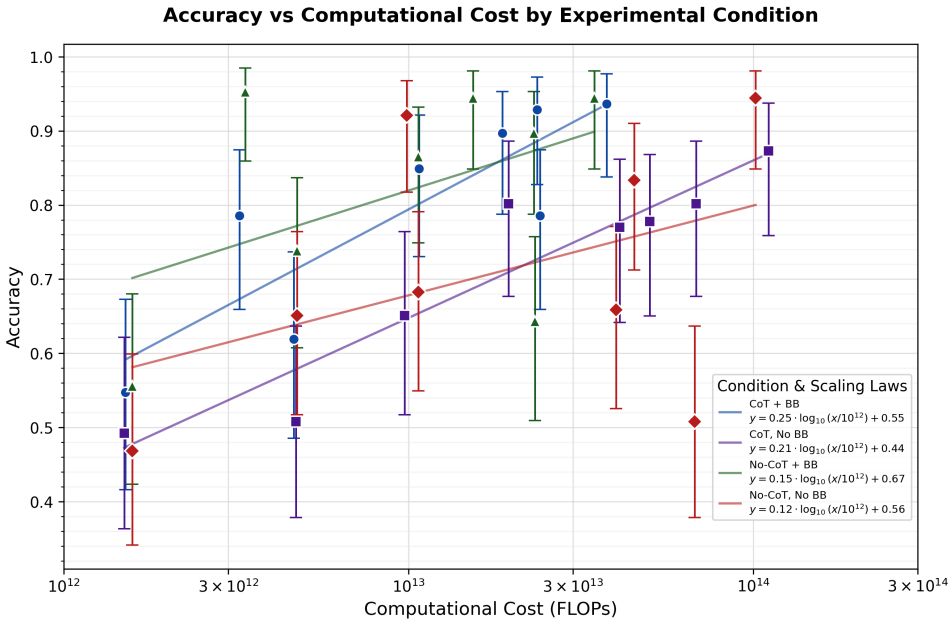

Figure 2: Plot of accuracy against computational cost for all models and experimental conditions. Error bars indicate the maximum and minimum values of the Wilson Score interval at the Z=3 (p =99.7%) significance level.

4. Conclusion

As educational institutions look to adopt AI, SLMs present a secure, transparent, and cost-effective alternative to large, cloud-based commercial models. However, effectively deploying SLMs requires architectural frameworks designed to compensate for their specific limitations. Our research demonstrates that when evaluating handwritten exams, providing SLMs with an entire double-page spread degrades their accuracy and unnecessarily increases token processing costs.

By implementing bounding boxes as a simple preprocessing step, we significantly reduced visual noise and token overhead. This intervention improved accuracy and reduced computational cost. We recommend that future research agendas focusing on SLMs in education should prioritise removing unnecessary visual clutter using techniques like bounding boxes, thereby optimising deployment of SLMs in learning analytics and automated grading.

Acknowledgments

This research was supported by the authors' respective institutions. We thank the organisers of the Australian Physics Olympiad for providing access to the anonymised student response dataset.

Declaration on Generative AI

During the preparation of this work, the author used Gemini 3-Pro to assist with literature review, formatting, preparing figures, drafting and proof reading. The author carefully reviewed and edited LLM generated content and take full responsibility for the publication's content.

References

[1] O. Henkel, L. Hills, A. Boxer, B. Roberts, Z. Levonian, Can Large Language Models Make the Grade? An Empirical Study Evaluating LLMs Ability To Mark Short Answer Questions in K-12 Education, in: Proceedings of the Eleventh ACM Conference on Learning @ Scale, L@S

'24, Association for Computing Machinery, New York, NY, USA, 2024, pp. 300–304. URL: https://dl.acm.org/doi/10.1145/3657604.3664693. doi:10.1145/3657604.3664693.

[2] G. Kortemeyer, Toward AI grading of student problem solutions in introductory physics: A feasibility study, Physical Review Physics Education Research 19 (2023) 020163.

[3] G. Kortemeyer, J. Nohl, Assessing confidence in AI-assisted grading of physics exams through psychometrics: An exploratory study, Physics Review Physics Education Research 21 (2025). doi:https://doi.org/10.1103/PhysRevPhysEducRes.21.010136.

[4] G. Kortemeyer, Performance of the pre-trained large language model GPT-4 on automated short answer grading, Discover Artificial Intelligence 4 (2024) 47.

[5] T. Liu, J. Chatain, G. Kortemeyer, M. Sachan, AI-assisted Automated Short Answer Grading of Handwritten University Level Mathematics Exams, 2024. URL: http://arxiv.org/abs/2408.11728. doi:10.48550/arXiv.2408.11728.

[6] R. Mok, F. Akhtar, L. Clare, C. Li, J. Ida, L. Ross, M. Campanelli, Using AI Large Language Models for Grading in Education: A Hands-On Test for Physics, 2024. URL: http://arxiv.org/abs/2411.13685. doi:10.48550/arXiv.2411.13685.

[7] Z. Chen, T. Wan, Achieving Human Level Partial Credit Grading of Written Responses to Physics Conceptual Question using GPT-3.5 with Only Prompt Engineering, Boston, 2024.

[8] Z. Chen, T. Wan, Grading explanations of problem-solving process and generating feedback using large language models at human-level accuracy, Physical Review Physics Education Research 21 (2025) 010126. URL: https://link.aps.org/doi/10.1103/PhysRevPhysEducRes.21.010126. doi:10.1103/PhysRevPhysEducRes.21.010126.

[9] L. McGinness, Overview of AI Grading of Physics Olympiad Exams, in: A. I. Cristea, E. Walker, Y. Lu, O. C. Santos, S. Isotani (Eds.), Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium, Blue Sky, and WideAIED, Springer Nature Switzerland, Cham, 2025, pp. 470–475. doi:10.1007/978-3-031-99261-2_55.

[10] P. Baumgartner, L. McGinness, Combining LLMs and Term Rewriting for Marking Algebraic Expressions in Physics Exams, in: KR Workshop: LLMs and KRR for Trustworthy AI, 2025. URL: https://openreview.net/forum?id=lir85TBz4V.

[11] C. D. Jaldi, A. Saini, S. Zhang, N. Schroeder, C. Shimizu, E. Ilkou, Small, Private Language Models as Teammates for Educational Assessment Design, 2026. URL: http://arxiv.org/abs/2605.15015. doi:10.48550/arXiv.2605.15015, arXiv:2605.15015 [cs.AI].

[12] L. McGinness, P. Baumgartner, Can large language models correctly interpret equations with errors?, Physical Review Physics Education Research 21 (2025) 020155. URL: https://link.aps.org/doi/10.1103/v8f8-s11v. doi:10.1103/v8f8-s11v.

[13] Z. Reza, A. Mazur, M. T. Dugdale, R. Ray-Chaudhuri, Small Models, Big Support: A Local LLM Framework for Educator-Centric Content Creation and Assessment with RAG and CAG, 2025. URL: http://arxiv.org/abs/2506.05925. doi:10.48550/arXiv.2506.05925, arXiv:2506.05925 [cs.CY].

[14] A. Kamath, J. Ferret, R. Dadashi, L. Hussenot, Gemma 3 Technical Report, 2025. URL: http://arxiv.org/abs/2503.19786. doi:10.48550/arXiv.2503.19786, arXiv:2503.19786 [cs.CL].

[15] A. Yang, B. Yang, Z. Qiu, Qwen2.5 Technical Report, 2025. URL: http://arxiv.org/abs/2412.15115. doi:10.48550/arXiv.2412.15115, arXiv:2412.15115 [cs.CL].

[16] P. K. A. Vasu, F. Faghri, C.-L. Li, O. Tuzel, H. Pouransari, FastVLM: Efficient Vision Encoding for Vision Language Models, 2025, pp. 19769–19780. URL: https://openaccess.thecvf.com/content/CVPR2025/html/Vasu_FastVLM_Efficient_Vision_Encoding_for_Vision_Language_Models_CVPR_2025_paper.html.

[17] J. Kaplan, S. McCandlish, T. Henighan, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, D. Amodei, Scaling Laws for Neural Language Models, 2020. URL: http://arxiv.org/abs/2001.08361. doi:10.48550/arXiv.2001.08361.

[18] S. C. Eimler, L. Erle, D. Flood, A. Haiman, L. Hackert, A. Helgert, L. McGinness, B. Yapici, The Environmental Cost of LLMs in AIED: Reporting and Practices, in: Lecture Notes in Computer Science (LNCS), Springer Nature Switzerland, 2026.

[19] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, D. Zhou, Chain-of-Thought Prompting Elicits Reasoning in Large Language Models, in: 36th Conference on Neural Information Processing Systems, NeurIPS, 2022.

[20] L. McGinness, P. Baumgartner, Large Language Models Imitate Logical Reasoning, but at what Cost?, 2025. doi:10.48550/arXiv.2509.12645.

[21] Bytez.com, R. Liu, J. Geng, A. J. Wu, I. Sucholutsky, T. Lombrozo, T. L. Griffiths, Mind Your Step (by Step): Chain-of-Thought can Reduce Performance on Tasks where Thinking Makes Humans Worse, 2025. URL: https://bytez.com/docs/icml/45714/paper.