Evaluating Vision–Language and Large Language Models for Automated Student Assessment in Indonesian Classrooms

Anonymous EMNLP submission

Abstract

Although vision language and large language models (VLM and LLM) offer promising opportunities for AI-driven educational assessment, their effectiveness in real-world classroom settings, particularly in underrepresented educational contexts, remains underexplored. In this study, we evaluated the performance of a state-of-the-art VLM and several LLMs on 646 handwritten exam responses from grade 4 students in six Indonesian schools, covering two subjects: Mathematics and English. These sheets contain more than 14K student answers that span multiple choice, short answer, and essay questions. Assessment tasks include grading these responses and generating personalized feedback. Our findings show that the VLM often struggles to accurately recognize student handwriting, leading to error propagation in downstream LLM grading. Nevertheless, LLM-generated feedback retains some utility, even when derived from imperfect input, although limitations in personalization and contextual relevance persist.

1 Introduction

Vision–language models (VLMs) (Liu et al., 2023, 2024b; Steiner et al., 2024) and large language models (LLMs) (Touvron et al., 2023a; Team, 2024; Team et al., 2024; OpenAI et al., 2024) have demonstrated impressive reasoning capabilities (Wang et al., 2023; Wei et al., 2022), including solving complex academic tasks such as university-level physics (Yeadon and Hardy, 2024) and competition-grade mathematics problems (Zhang et al., 2024). These advancements have driven growing interest in applying such models to education. Common areas of application include automated grading (Chiang et al., 2024), teaching support (Hu et al., 2025), feedback generation Morris et al. (2023), and content creation (Westerlund and Shcherbakov, 2024).

However, most VLM and LLM-based educational tools have been developed with English-speaking contexts in mind Lee and Zhai (2025); Yancey et al. (2023), limiting their relevance and usability in non-English-speaking regions, particularly in rural areas in Indonesia. Ensuring sociocultural relevance is essential: effective deployment requires adaptation to local curricula, languages, and cultural norms, rather than relying on a one-size-fits-all approach. Moreover, the shortage of qualified teachers in rural areas highlights the importance of prioritizing AI integration in underserved regions, rather than concentrating development efforts solely in high-resource, Global North contexts (Jin et al., 2025; Kristiawan et al., 2024).

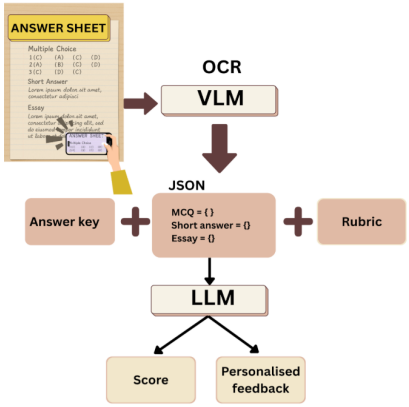

Figure 1: AI-powered assessment using VLM and LLM.

In this study, we address the contextual challenges of applying AI-powered assessment tools in non-English speaking and under-resourced settings by collecting real-world student assessment data from primary schools in the form of handwritten responses. This design is motivated by two practical considerations. First, many schools, especially in rural areas, lack consistent access to digital devices, highlighting the need for AI systems that function effectively in low-tech environments. Second, using handwritten responses helps reduce the risk of academic dishonesty, such as students who rely on AI tools to generate answers. Assessments were conducted in Indonesian for the mathematics subject, while the responses to the English subject

were written in English, reflecting the language of instruction for each subject.

Our contributions are as follows: (1) We release a dataset of 646 handwritten student answer sheets (with over 14K answers) collected from six primary schools in Indonesia—three from rural areas and three from urban areas. The assessments cover Grade 4 Mathematics and English, with questions and scoring guidelines developed by experienced teachers. All student responses were manually transcribed and graded by professional teachers.[1] (2) We introduce a multimodal pipeline that integrates vision–language models (VLMs) and large language models (LLMs), as illustrated in Figure 1. We compare several state-of-the-art models for grading student answers and find that GPT-4o with vision input achieves the highest accuracy and feedback quality. (3) We conduct a manual evaluation of LLM-generated feedback in Indonesian and find that, even when based on imperfect input (e.g., OCR errors), the feedback tends to be clear and factually correct. However, personalization and helpfulness remain notable areas of concern.

2 Related Work

Previous studies have investigated the use of LLMs as graders for student assignments and exams. For example, Chiang et al. (2024) used GPT-4 to automatically grade 1,028 student essays in a university-level course titled *Introduction to Generative AI*. Their findings suggest that LLM-based graders were generally well accepted by students; however, the models occasionally did not follow the grading rubric. In a related study, Yancey et al. (2023) used GPT-3.5 and GPT-4 to score essays in a high-stakes English proficiency test, demonstrating that LLM-generated scores can achieve high agreement with human raters.

Stahl et al. (2024) used Mistral (Jiang et al., 2023) and LLaMA-2 (Touvron et al., 2023b) to assess English student essays and generate feedback, finding that scoring accuracy had limited influence on student's perceived usefulness of the feedback. Similarly, Morris et al. (2023) applied a Longformer-based language model (Botarleanu et al., 2022) to generate formative feedback on student-written summaries of English textbooks.

Unlike these prior studies, our work focuses on handwritten responses from grade 4 primary school students in Indonesia, covering both English and mathematics. We also evaluate a complete multimodal pipeline that integrates a VLM for handwriting recognition and LLMs for grading and feedback generation—introducing new challenges related to noisy input, multilingual content, and real-world constraints in low-tech, underrepresented classroom settings.

3 Dataset Construction

Assessment Design We developed assessment instruments for grade 4 primary school students in two subjects: Mathematics and English. The items were designed from scratch based on a thorough analysis of the national curriculum and corresponding learning objectives. Each subject assessment consisted of 10 multiple-choice questions (MCQs), 10 short-answer questions, and 2 essay questions. All items were created by experienced senior subject teachers—an English teacher and a Math teacher—each with over 10 years of classroom experience and a Master's degree in Education. In addition to writing the assessment items, these teachers developed detailed scoring rubrics for the short-answer and essay questions, as well as answer keys for the MCQs. Standardized answer sheets were also prepared to collect student responses.

Data Collection Data collection was carried out in six primary schools, evenly divided between rural (Sumatra and Nusa Tenggara Islands) and urban (Java Island) settings. Each classroom included approximately 20 to 30 students. For both subjects, students followed a structured sequence consisting of a pre-test, lesson, and post-test. Students had up to 30 minutes to complete their answers on a standardized answer sheet.

In total, we collected 646 handwritten answer sheets from these assessments, comprising both pre-tests and post-tests. Of these, 414 were collected from urban schools and 232 from rural schools. The disparity in sample size between urban and rural areas is primarily due to larger class sizes typically found in urban schools compared to their rural counterparts.

4 Experiment

Overall Pipeline Figure 1 illustrates our pipeline, which begins with a vision–language model (VLM)

[1] To ensure ethical use and protect student privacy, all personally identifiable information (e.g., student names, grade levels, and school names) has been removed.

that performs optical character recognition (OCR) to extract handwritten student responses from scanned answer sheets. The extracted text is then structured into a JSON format and passed to a large language model (LLM), along with the answer key and a teacher-defined rubric. For multiple-choice questions, we apply string matching. For short-answer and essay questions, we run the LLM separately for each question, providing the student's response, the corresponding answer key, and the assessment rubric. To generate personalized feedback, we provide the LLM with all of the student's responses, the answer key, the assigned weights, and the rubric.

Model For OCR, we use GPT-4o (OpenAI et al., 2024), alongside a gold-standard transcription manually parsed by teachers. For automatic scoring, we compare the performance of GPT-4o, Llama-3.1-Instruct (70B) (Touvron et al., 2023b), Qwen2.5-Instruct (72B) (Team, 2024), and Deepseek-Chat (671B) (Liu et al., 2024a). For generating personalized feedback, we rely on the scoring results produced by GPT-4o and generate two versions of feedback using GPT-4o and Deepseek-Chat. All prompts and decoding hyperparameters used are provided in the Appendix.

Evaluation Each answer sheet image was manually transcribed and scored by professional teachers. We compared the LLM-generated scores against these gold-standard scores across three question types: multiple-choice, short-answer, and essay, using mean absolute error (MAE) as the evaluation metric. For personalized feedback, we conducted a manual evaluation covering four aspects—Correctness, Personalization, Clarity, and Educational Value/Helpfulness—rated on a 1–5 scale, where 1 indicates the lowest quality.[2]

5 Result and Analysis

Main Result Table 1 presents the performance of the LLMs selected in three types of questions: multiple choice, short answer, and essay. When using GPT-4o to extract student responses via OCR, we observe that most model-generated scores are generally competitive. Among them, GPT-4o produces scores that align most closely with human grading for essay questions, achieving the lowest MAE in both English (5.6) and Math (1.5). In contrast, LLaMA-3.1–70B and Qwen-2.5–72B are less reliable, with scores deviating more significantly from human judgments. Short-answer questions remain the most challenging to evaluate: even the best performing model in this category, LaMA-3.1-7B for Math, still shows a relatively high MAE of 10.6, indicating a notable gap from human-level accuracy.

However, the results differ when human effort is involved in the OCR task. Most scores become better overall, with Deepseek-chat and GPT-4o emerging as the top-performing models. Deepseek-chat shows strong performance in English (MAE of 4.4 for short answers and 1.5 for essays), while GPT-4o performs best in Math, with only a 2.9 difference in short answers and 5.7 in essays. It is worth noting that MCQ scores remain at 0, as basic string matching is sufficient due to the exact nature of the answers. The impact of OCR performance on LLM scoring is further discussed in Section 5.

Human Evaluation on Personalised Feedback Table 2 presents the results of a human evaluation on personalized feedback quality, rated by expert teachers across four dimensions: Correctness, Per-

Model	English				Math			
	M	S	E	Total	M	S	E	Total
OCR by GPT4o								
GPT4o	**2.8**	14.6	**5.6**	**11.7**	**2.3**	16.3	**1.5**	8.2
Llama 3.1 (70B)	**2.8**	18.7	9.3	14.5	**2.3**	**10.6**	27.5	**2.2**
Qwen2.5 (72B)	**2.8**	14.9	16.6	14.7	**2.3**	19.1	5.8	7.1
Deepseek (671B)	**2.8**	**12.6**	9.8	11.9	**2.3**	22.8	6.7	8.1
OCR by Human								
GPT4o	**0.0**	9.2	2.7	7.9	**0.0**	**2.9**	**5.7**	1.5
Llama 3.1 (70B)	**0.0**	14.4	2.3	11.6	**0.0**	9.8	19.1	10.3
Qwen2.5 (72B)	**0.0**	8.4	3.8	9.2	**0.0**	5.5	8.7	3.3
Deepseek (671B)	**0.0**	**4.4**	**1.5**	**6.8**	**0.0**	5.9	8.5	**0.8**

Table 1: Mean absolute error (MAE) for English and Math, calculated separately for multiple-choice (M), short-answer (S), essay (E), and the total score. Lower values indicate better performance; bolded numbers represent the best results. Scores for each component range from 0 to 100.

Model	Correctness	Personalization	Clarity	Helpfulness
English				
GPT-4o	**4.00**	**3.96**	3.64	3.60
Deepseek	3.96	3.88	**4.04**	**3.96**
Math				
GPT-4o	3.84	**3.72**	3.92	**3.68**
Deepseek	**3.88**	2.96	**4.00**	2.92

Table 2: Human evaluation by expert teachers on personalized feedback, using a rating scale from 1 to 5, where 1 indicates the lowest score.

[2]This evaluation was carried out by an experienced educator with a Master's degree in teaching. The evaluation guidelines and definitions for each aspect are provided in the Appendix.

Model	English				Math			
	M	S	E	Total	M	S	E	Total
Urban								
GPT4o	**0.0**	2.4	7.2	0.8	**0.0**	5.8	**7.6**	2.4
Llama 3.1 (70B)	**0.0**	7.7	2.9	2.7	**0.0**	10.3	30.0	10.4
Qwen2.5 (72B)	**0.0**	1.9	**1.3**	**0.5**	**0.0**	7.6	10.7	3.9
Deepseek (671B)	**0.0**	**1.3**	3.5	1.5	**0.0**	**5.6**	9.9	**1.0**
Rural								
GPT4o	**0.0**	21.2	**5.2**	23.1	**0.0**	2.5	**2.2**	**0.3**
Llama 3.1 (70B)	**0.0**	26.1	11.4	26.9	**0.0**	8.8	23.1	9.7
Qwen2.5 (72B)	**0.0**	19.8	12.5	24.3	**0.0**	**1.7**	5.0	2.1
Deepseek (671B)	**0.0**	**14.2**	10.1	**21.2**	**0.0**	6.4	5.9	0.6

Table 3: Analysis of mean absolute errors (MAE) for English and Math across urban and rural settings, calculated separately for multiple-choice (M), short-answer (S), essay (E), and total scores. The OCR results used in this analysis were obtained through **human transcription**. Lower values indicate better performance; bolded values represent the best results. Each component is scored on a 0–100 scale.

sonalization, Clarity, and Helpfulness (scale 1–5, with scores below 3 considered poor). For English, GPT-4o slightly outperforms Deepseek in correctness and personalization, while Deepseek leads in clarity and helpfulness. In Math, Deepseek shows strong clarity and correctness but performs poorly in personalization and helpfulness, with both scores falling below 3. GPT-4o, on the other hand, maintains more balanced performance across all dimensions.

Urban vs. Rural Performance Analysis Given the significant educational disparities between rural and urban areas, we evaluated the performance of the model in these two settings. To isolate the analysis of LLM scoring capabilities, we use only the human-transcribed OCR results, eliminating recognition errors.

Table 3 presents the MAE scores for English and Math, separated by question type: multiple choice (M), short answer (S), essay (E), and total scores. The results indicate that English MAEs are generally higher in rural settings than in urban settings across all models. For example, GPT-4o achieves a total MAE of only 0.8 in urban English, but this rises sharply to 23.1 in the rural setting. This discrepancy suggests that LLMs may struggle more in interpreting free-form responses from rural students, possibly due to variations in writing style and grammar. In contrast, MAEs for Math tend to be slightly lower in rural areas, although the differences are less pronounced. This may be attributed to the nature of Math questions, which often involve numerical reasoning and have more deterministic answers, reducing ambiguity in scoring.

Area	English			Math		
	EM(M)	EM(S)	RL(E)	EM(M)	EM(S)	RL(E)
Urban	**82.1**	**67.1**	**60.3**	62.3	23.3	21.0
Rural	71.7	61.8	60.1	**62.5**	**27.9**	**24.8**
All	78.5	65.3	60.2	62.4	24.9	22.3

Table 4: OCR-based performance (GPT-4o) across Urban, Rural, and All settings for English and Math: EM = exact match, RL = ROUGE-L F1, MCQ = multiple choice.

OCR Performance Analysis Given the differences in MAE between the GPT-4o OCR outputs and human transcription shown in Table 1, we further analyze the OCR performance of GPT-4o and evaluate the extent to which recognition errors propagate to the subsequent scoring. For this analysis, we use exact string matching to assess accuracy on multiple choice and short answer questions, and compute ROUGE-L (Lin, 2004) scores to compare GPT-4o and human transcriptions for essay questions.

Table 4 shows that the OCR performance is generally higher for English than for Math. Within English, responses from urban students yield higher exact match and ROUGE-L scores compared to those from rural students, possibly due to differences in handwriting clarity or writing conventions. For Math, the OCR accuracy is overall lower than that of English, but the performance gap between urban and rural settings is less pronounced. This suggests that while English responses may be more affected by region-specific handwriting variability, Math responses, often more structured and numerical, are comparatively stable across regions.

6 Conclusion

In this work, we present a real-world implementation of vision–language model (VLM) and large language models (LLMs) for student assessment in underrepresented regions—specifically, rural and urban areas of Indonesia—focusing on primary school subjects in Math and English. Our results show that GPT-4o and Deepseek (671B) perform competitively in matching teacher-assigned scores across multiple-choice, short-answer, and essay formats. For personalized feedback generation, manual evaluation indicates that Deepseek outperforms GPT-4o in terms of quality and relevance. We hope that this work encourages greater research attention towards educational applications of AI in low-resource and underserved contexts.

Limitations

While this study provides valuable insights into the use of vision-language and large language models (VLMs and LLMs) for automated assessment in multilingual, low-resource contexts, several limitations should be acknowledged:

Educational Scope The study was conducted exclusively in Indonesian public elementary schools, specifically in Grade 4 classrooms following the national curriculum (*Kurikulum Merdeka*). It focused on two subject areas: Mathematics (covering the introductory chapter on fractions) and English (focusing on the topic of parts of the house). As such, the findings may not be generalizable to other subjects, grade levels, or curricula. Geographically, the research was limited to three provinces—West Java (Java Island), West Nusa Tenggara (Lombok Island), and West Sumatra (Sumatra Island)-–which, while diverse, may not fully represent the broader variation in educational contexts across Indonesia or other countries.

Models The models used in our evaluations include OpenAI's GPT-4o, Meta's LLaMA 3.1–70B Instruct, Qwen 2.5–VL–72B Instruct, and DeepSeek Chat. While these models represent the current state of the art, their training data and evaluation strategies are primarily optimized for English and other globally dominant contexts. As a result, they may struggle to fully capture the nuances of student responses written in Bahasa Indonesia.

Ethics Statement

This study strictly adheres to ethical research practices in AI and education:

- All student answer sheets were anonymized prior to analysis. Identifying information, including names, school names, and class identifiers, was removed to protect student privacy and comply with ethical guidelines for research involving minors.
- Written informed consent was obtained from school administrators and participating teachers. Participation in the study was voluntary, and students were not penalized for opting out.
- The inclusion of both urban and rural schools was an intentional decision to ensure representation across socio-economic and educational divides. However, we recognize that the deployment of AI tools in such settings must be approached cautiously to avoid reinforcing existing inequalities. This study advocates for equitable development, localization, and participatory design of AI tools in education, particularly when applied in under-resourced areas.
- To mitigate risks associated with overreliance on AI outputs, all AI-generated scores and feedback were reviewed by experienced teachers. We emphasize that AI should augment—not replace—human judgment in educational assessment, especially when dealing with young learners.

References

Robert-Mihai Botarleanu, Mihai Dascalu, Laura K Allen, Scott Andrew Crossley, and Danielle S McNamara. 2022. Multitask summary scoring with longformers. In *International Conference on Artificial Intelligence in Education*, pages 756–761. Springer.

Cheng-Han Chiang, Wei-Chih Chen, Chun-Yi Kuan, Chienchou Yang, and Hung-Yi Lee. 2024. Large language model as an assignment evaluator: Insights, feedback, and challenges in a 1000+ student course. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2489–2513.

Bihao Hu, Jiayi Zhu, Yiying Pei, and Xiaoqing Gu. 2025. Exploring the potential of llm to enhance teaching plans through teaching simulation. *npj Science of Learning*, 10(1):7.

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

Yueqiao Jin, Lixiang Yan, Vanessa Echeverria, Dragan Gašević, and Roberto Martinez-Maldonado. 2025. Generative ai in higher education: A global perspective of institutional adoption policies and guidelines. *Computers and Education: Artificial Intelligence*, 8:100348.

Dana Kristiawan, Khaliq Bashar, and Dian Arief Pradana. 2024. Artificial intelligence in english language learning: A systematic review of ai tools, applications, and pedagogical outcomes. *The Art of Teaching English as a Foreign Language (TATEFL)*, 5(2):207–218.

Gyeonggeon Lee and Xiaoming Zhai. 2025. Realizing visual question answering for education: Gpt-4v as a multimodal ai. *TechTrends*, pages 1–17.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. 2024b. Vmamba: Visual state space model. *Advances in neural information processing systems*, 37:103031–103063.

Wesley Morris, Scott Crossley, Langdon Holmes, Chaohua Ou, Danielle McNamara, and Mihai Dascalu. 2023. Using large language models to provide formative feedback in intelligent textbooks. In *International Conference on Artificial Intelligence in Education*, pages 484–489. Springer.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Pe-

ter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko, Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. 2024. Gpt-4o system card.

Maja Stahl, Leon Biermann, Andreas Nehring, and Henning Wachsmuth. 2024. Exploring llm prompting strategies for joint essay scoring and feedback generation. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 283–298.

Andreas Steiner, André Susano Pinto, Michael Tschannen, Daniel Keysers, Xiao Wang, Yonatan Bitton, Alexey Gritsenko, Matthias Minderer, Anthony Sherbondy, Shangbang Long, et al. 2024. Paligemma 2: A family of versatile vlms for transfer. *arXiv preprint arXiv:2412.03555*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. 2024. Gemma 2: Improving open language models at a practical size.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Boshi Wang, Xiang Yue, and Huan Sun. 2023. Can chatgpt defend its belief in truth? evaluating llm reasoning via debate. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11865–11881.

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

Magnus Westerlund and Andrey Shcherbakov. 2024. Llm integration in workbook design for teaching coding subjects. In *International Conference on Smart Technologies & Education*, pages 77–85. Springer.

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.

Will Yeadon and Tom Hardy. 2024. The impact of ai in physics education: a comprehensive review from gcse to university levels. *Physics Education*, 59(2):025010.

Di Zhang, Xiaoshui Huang, Dongzhan Zhou, Yuqiang Li, and Wanli Ouyang. 2024. Accessing gpt-4 level mathematical olympiad solutions via monte carlo tree self-refine with llama-3 8b. *arXiv preprint arXiv:2406.07394*.

A Hyperarameter Setup

We use the following default hyperparameters: temperature = 1.0, top-p = 1.0, and top-k = 1.0 for all tasks, including OCR of student papers, scoring, and generating feedback. The max_tokens parameter is also set to its default to allow the model to generate output without restrictions.

B Prompts List

Figures 2, 3, and 4 show the prompts we use to generate outputs for the OCR task, score student answers, and provide feedback based on the student's assignment performance.

Prompt for reading the image (OCR)

This is an image of an answer sheet with texts written in either English or Indonesian. Please extract all answers from the image. Adjust the numbering in your response to match the actual number of questions on the answer sheet. Use the following JSON format in your output, and do not output anything else.

```
{
    'Nama': <value>,
    'Kelas': <value>,
    'PILIHAN GANDA': {
        '1': <value>,
        '2': <value>,
        // Adjust numbering based on the answer sheet},
    'ISIAN': {
        '1': <value>,
        '2': <value>,
        // Adjust numbering based on the answer sheet},
    'ESSAY': {
        '1': <value>,
        '2': <value>,
        // Adjust numbering based on the answer sheet}, }
```

Figure 2: Prompt for reading the image (OCR) using LLM

Prompt for scoring

The maximum score for this question is {max_score}. Please follow this marking criteria when deciding the score for the student's answer

{marking_criteria}

Student answer:

{student_answer}

Answer key:

{gold_answer}

What is the appropriate score for the student in a range of 0 and {max_score}? Please only output the score in your response!'

Figure 3: Prompt for scoring using LLM

Prompt for generating the feedback

Write in Indonesian a personalised feedback (less than 8 sentences) for a student {student_name} based on the evaluation results over his/her exam answer.
Please use this JSON data by focusing on obtained_score and learning_objective.

{detailed_feedback}

Figure 4: Prompt for generating the feedback using LLM

C Human Evaluation Guideline on Personalised Feedback

We evaluate the quality of personalized feedback along four dimensions using a 1–5 rating scale, where 1 indicates the lowest quality and 5 indicates the highest. The four dimensions are **Correctness**, **Personalization**, **Clarity**, and **Educational Value / Helpfulness**. *Correctness* assesses whether the feedback is factually accurate based on the student's response, the answer key, and the rubric. *Personalization* measures how well the feedback is tailored to the student's specific answer, including whether it addresses actual strengths, weaknesses, or errors rather than offering generic comments. *Clarity* evaluates whether the feedback is easy to understand, well-structured, and communicated in an age-appropriate and supportive tone. *Educational Value / Helpfulness* considers the extent to which the feedback supports learning and encourages the student to reflect and improve. Evaluators are instructed to use these criteria consistently when assigning scores.