L(V)LMs Compromising the Integrity of in-Person Exams: An Evaluation Utilizing Smart Glasses and Computer Vision

Rupert Urbanski[a] and Ralf Peters[b]
Institut für Wirtschaftsinformatik und Operations Research, Martin-Luther-Universität Halle-Wittenberg, Universitätsring 3, 06108 Halle (Saale), Germany

Keywords: Large Language Models, Large Vision-Language Models, LLM, LVLM, Exam, Threat, Integrity, Smart Glasses, Computer Vision, CV, Cheating, Smart Devices, Benchmark.

Abstract: This article assesses the threat of LVLMs and smart glasses that interface with them towards the integrity of in-person exams proposes and utilizes a new specifically developed benchmark based on standardized exam questions. While performance decreases under image degradation are being demonstrated, it still highlights the high accuracy of publicly available models when answering exam questions, even under less-than-optimal conditions, which showcases the need for researching more robust exams. Additionally, approaches to developing benchmarks whose performance translates better to real-life scenarios are being demonstrated, along with quantifying expected performance detriments when moving from synthetic benchmarks under ideal conditions to similar practical applications.

1 INTRODUCTION

With the increasing capabilities of newer machine learning models, which can process more diverse data to achieve more generalizability, developing representative benchmarks to measure their overall performance becomes more complex. Furthermore, it is of interest to make these measurements understandable to humans and allow comparison in common terms. This is why exam questions have been established to display and measure said capabilities. (Cobbe et al., 2021; Hendrycks et al., 2021; OpenAI, 2023; Shen et al., 2021)

Whereas these papers propose a new dataset for benchmarking or employ these as targets to increase scores on them for newly developed models, not much regard is being paid to ensuring exam integrity.

While online exams are inherently more susceptible to cheating using electronic devices, this issue isn't limited to them. This is especially due to the current simultaneous rise of simple smart glasses, such as those by a collaboration of Ray-Ban and Meta (Ray-Ban, 2024) or Spectacles by Snap Inc. Lacking visual output through display makes them practically indistinguishable from their non-smart counterparts to the untrained eye. Having small, integrated, high-resolution cameras and interfaces for communicating via the Internet allows easy input to Large (Vision-)Language Models (LLMs/LVLMs), which have proven to perform well on known standardized exams. (OpenAI, 2023) This closes a gap in previous research on using smart devices, which require less concealable input or collaboration. (Heya, Serwadda, Griswold-Steiner, & Matovu, 2021; Nyamawe & Mtonyole, 2014; Wong, Yang, Riecke, Cramer, & Neustaedter, 2017)

Since these smart glasses are also openly available and have gathered a broader public audience than earlier developments such as Google Glasses, especially in modern social media, using them to cheat in in-person exams could become a widespread issue.

Standardized exams can still pose a challenge to L(V)LMs since they involve reasoning, especially in the form of selecting the correct answer in a mix of single-choice (SCQ) and free response questions (FRQ). They also consist of tasks in heterogeneous disciplines, such as reading comprehension, which involves sentence completion and interpreting formulas and graphs. With the addition of having

[a] https://orcid.org/0009-0000-1512-6392
[b] https://orcid.org/0009-0007-9168-0126

Urbanski, R. and Peters, R.
L(V)LMs Compromising the Integrity of in-Person Exams: An Evaluation Utilizing Smart Glasses and Computer Vision.
DOI: 10.5220/0013206100003932

In *Proceedings of the 17th International Conference on Computer Supported Education (CSEDU 2025) - Volume 2*, pages 43-50
ISBN: 978-989-758-746-7; ISSN: 2184-5026

pictures taken by smart glasses as input, which suffer from noise, skewed perspective, and other degrading factors, solving exam questions via L(V)LMs is far from trivial.

The idea to show the feasibility and evaluate the current threat has been explored and established conceptually in a previous talk session by the authors of this paper. (Rupert Urbanski and Ralf Peters, 2024)

Thus, the contributions of this paper are threefold: First, it brings the aforementioned concepts into practice and evaluates the situation systematically; second, it moves from known synthetic benchmarks to a real-life scenario and explores concepts for benchmarking with results that translate better to performance in practice; and third, with assessing and showcasing the compromised integrity of standardized exams it also opens up the debate on how to make said exams more robust to LLMs, which could be similar to research field of CAPTCHAs.

This article is structured as follows: Section 1 is this introduction. Section 2 is devoted to related work. Section 3 describes the developed dataset. Section 4 introduces the study layout, including the applied methodology, and contains evaluations of the study results.

2 RELATED WORK

These have previously been surveyed and can be categorized by various criteria. The given task is similar to known Natural Language Processing (NLP) tasks such as Question Answering (QA).

QA usually involves multiple steps: classifying the question, retrieving relevant information, and extracting an answer. Allam and Haggag have surveyed traditional ML approaches to these core challenges in the research field. These approaches often include explicit modeling specific to subdomains for these subtasks. (Allam & Haggag, 2012)

LLMs and the extension LVLMs allow a more generalized approach to problems, utilizing a more extensive knowledge base. Comparisons between their capabilities utilize a wide array of benchmarks for L(V)LMs, some of which can be categorized as QA-focussed benchmarks. (Chang et al., 2023; Guo et al., 2023; Zhou, Guo, Wang, Chang, & Wu, 2024)

Furthermore, the task Visual Questions Answering (VQA), which sits between Computer Vision (CV) and QA, has also been surveyed separately. (Wu et al., 2017) A noteworthy example in this field is VQAv2, which hints at models learning language priors and not actually answering based on the image. (Goyal et al., 2019)

VQA differs from the given task of answering exam questions in image form since it usually involves a separate prompt, which is given in addition to an image and only determined at run time. The images used in VQA tasks also typically do not include or focus on textual tasks but on the content or details in the image. (Wu et al., 2017)

The given task differs in this regard and is also similar to Optical Character Recognition (OCR), another known task in CV, that involves recognizing and extracting textual information from images and has been surveyed for LVLMs. (Xu et al., 2023).

Previous articles in the OCR research field have addressed noisy or degraded image data and the tolerance of OCR engines to those issues in the past. These give insights into artificially generating realistic noise and possible performance impacts. (Baird, 2007; Guyon, Haralick, Hull, & Phillips, 2000)

Leveraging LVLMs for more specialized and integrated OCR tasks and integrating further interference steps as in VQA have been explored in multiple benchmarks such as OCRBench (Liu et al., 2023) and DocVQA (Mathew, Karatzas, & Jawahar, 2020). These involve reading comprehension but don't regard sentence completion as in standardized exams and don't address image degradation systematically. While HellaSwag (Zellers, Holtzman, Bisk, Farhadi, & Choi, 2019) does feature sentence completion, these are aimed towards commonsense, not reading comprehension on an academic level. ChartQA (Masry, Long, Tan, Joty, & Hoque, 2022) benchmarks VQA capabilities for charts, which is a subtask in standardized exams, in addition to calculations. The latter is more thoroughly addressed in MathVista (Lu et al., 2023), which also involves selecting a correct answer to single choice question and consists of various preexisting and new datasets.

The field of Arithmetic Reasoning involves more interference based on text-based math problems, known benchmarks like GSM8K (Cobbe et al., 2021) and MultiArith (Roy & Roth, 2016). However, they are typically designed based on lower education levels than standardized exams and don't involve any image processing.

In the field of LLMs, standardized exams have been utilized to gauge performance in relation to humans. Most commonly known and State-of-the-Art, OpenAI employs various university entrance and advanced placement level exams for this regard, giving the exercises to the models in pre-processed text form. (OpenAI, 2023).

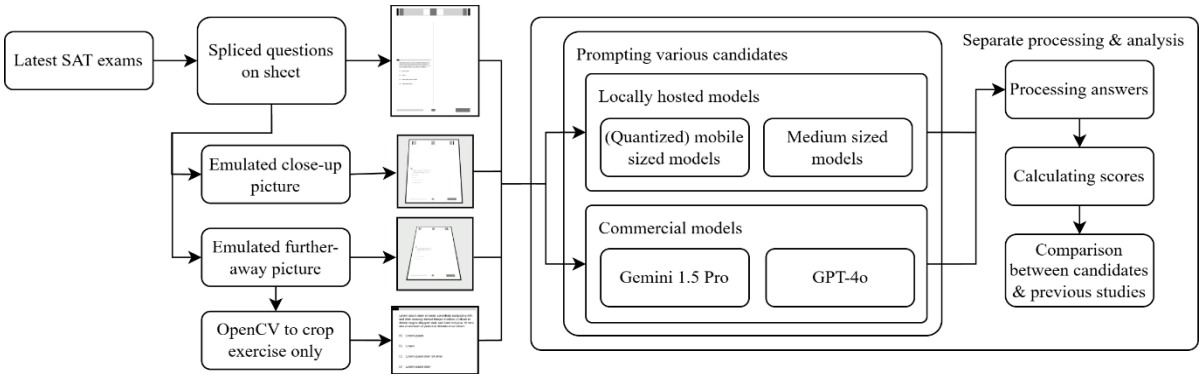

Figure 1: Overall study layout.

OpenAI also utilizes Multi-shot-prompting, which involves giving the model examples, including answers similar to, e.g., the current exercise to be answered. While this isn't realistic for the given task, it hints towards optimizing prompts to increase model performance. One way to do this while maintaining Zero-shot-prompting, which only involves giving the current to be-answered exercise, is utilizing chain of thought (CoT) prompting, as Kojima, Gu, Reid, Matsuo & Iwasawa proposed. Their paper also includes an evaluation that shows that Zero-shot-CoT-prompting has proven effective in improving performance in the aforementioned Arithmetic Reasoning benchmarks. (Kojima, Gu, Reid, Matsuo, & Iwasawa, 2022)

These previous works give a rough estimate of what to expect from a model but do not integrate all the necessary challenges to represent performance on a whole real-life task such as the one addressed in this paper.

However, they also give an overview of possible problems, along with appropriate approaches to these and promising models to be evaluated for the given task, as well as evaluation metrics commonly used in the field.

3 METHODOLOGY AND EXPERIMENTS

Four scenarios have been modeled to evaluate the reach and various approaches to the task. Figure 1 shows an overview of the study layout.

The basis for these scenarios is a hand-crafted task set that has been derived from publicly available SAT exams from the previous years, similar to OpenAIs used dataset for evaluation in their technical report (OpenAI, 2023), which also ensures comparable results.

To get answers to the expected task when prompting, these exams have been modelled to only contain a single exercise per page in opposition to the public available pagers but still maintain the rest of the page context.

This eliminates the additional challenge of finding the requested exercise on a page among multiple exercises and thus also lowers the possibility of model confusion but is still realistic since it can be done by covering parts of the page in practice

This results in 720 images/exercises per dataset, which consist of 396 reading comprehension tasks and 324 math problems. The reading comprehension tasks include sentence completion tasks, along with 13 tasks that involve data comprehension in tables. Furthermore, some tasks refer to underlined sections, which isn't part of most VQA benchmarks. All the reading tasks are single choice with four lettered options from A to D (SCQ). 240 math problems have the same format, and 84 are to be answered freely (FRQ). 14 of these tasks include tables; 50 involve graphs or other figures. An overview of these proportions is shown in Table 1.

Table 1: Composition of task set.

task		format		including	
math problems	324	FRQ	84	tables	1
				figures	8
		SCQ	240	tables	13
				figure	42
reading comprehension	396	SCQ	396	figures	30
				tables	13

The first scenario consists of just spliced pages with isolated tasks. The following three subsections show the further processing to derive the datasets for the other scenarios.

Table 2: Accuracy on cropped exercises with selected models highlighted (grey)

Model	Math/FRQ	Math/SCQ	Math/Overall	Reading	Overall
google/paligemma-3b-mix-224	3,57%	17,50%	13,89%	9,09%	11,25%
google/paligemma-3b-mix-448	0,00%	23,33%	17,28%	25,76%	21,94%
llava-hf/llava-onevision-qwen2-0.5b-si-hf	7,14%	22,50%	18,52%	25,76%	22,50%
llava-hf/llava-onevision-qwen2-7b-si-hf	30,95%	39,17%	37,04%	37,37%	37,22%
meta-llama/Llama-3.2-11B-Vision-Instruct	70,24%	52,08%	56,79%	27,78%	40,83%
llava-hf/llava-v1.6-vicuna-13b-hf	4,76%	29,17%	22,84%	30,05%	26,81%
llava-hf/llama3-llava-next-8b-hf	21,43%	30,83%	28,40%	24,49%	26,25%
llava-hf/llava-v1.6-vicuna-7b-hf	10,71%	25,00%	21,30%	30,05%	26,11%
google/gemini-1.5-pro-002	94,05%	92,08%	92,59%	78,03%	84,58%
openai/gpt-4o	88,10%	93,75%	92,28%	83,08%	87,22%

3.1 Emulated Further-away Pictures

To emulate a close-to-realistic usage scenario, images of an exam page have been taken and analysed to get parameters to degrade image quality accordingly using batch processing.

This involves perspective-based skewing, which is close to a 3D rotation of 75° around the y-axis, applying a Gaussian Blur in a radius of 3 pixels, and adding 5% Gaussian Noise to the image.

These parameters have been determined utilizing grid search and utilizing perceptual image metrics proposed by Yee (Yee, 2004) and Zauner (C. Zauner, 2010) to determine image similarity. Further comparative analysis to confirm the results has also been done manually.

3.2 Emulated Close-up Picture

To systematically evaluate the impact of this image degradation, the same methods as in 3.1 have been applied with reduced (around half) parameter values.

This involves perspective-based skewing which is close to a 3D rotation of 32° around the y-axis, applying a Gaussian Blur in a radius of 1 pixel and adding 3% Gaussian Noise to the image.

3.3 Cropped Exercises

The images of the emulated further-away pictures have been processed further to establish a more baseline performance of the evaluated LVLMs under more ideal conditions, enable model selection, and emulate a more sophisticated approach, e.g., using a specialized app for solving exercises.

Specifically, OpenCV has been utilized to de-skew the pages, reduce noise, sharpen the image, determine the exercises' bounding boxes, and crop the images accordingly.

4 EVALUATION

Six leading and freely available models were selected from the Hugging Face leaderboards for the benchmarks mentioned in section 2. Due to hardware constraints, only models with up to 11 billion parameters were considered.

This selection contains Llama 3.2 Vision by Meta, which has been tuned using an instruction dataset (Meta, 2024) and multiple models from the LLaVA-OneVision family, which utilize various preexisting LLMs as language backbone along with a common vision encoder (Li et al., 2024) and have been adapted for usage via Hugging Face's transformers library.

Additionally, three smaller models have been selected that can run locally on mobile devices to evaluate feasibility as a mobile application without involving external computing services. These include two models by Google's PaliGemma family, which have been trained on 224x224 and 448x448 pixel-sized images, respectively. These models have shown good results for VQA tasks but are not intended for conversational use. (Beyer et al., 2024) The third is a smaller version of a LLaVa-OneVision model, which uses Qwen2 as a base LLM with 0.5 billion parameters. (Li et al., 2024)

Two commercial State-of-the-art models, OpenAI's GPT-4o (OpenAI, 2023) and Google's Gemini 1.5 Pro-002 (Georgiev et al., 2024), which are available as cloud services, also get assessed.

All models have been prompted utilizing the same Zero-shot-CoT prompts for single-choice and free-response exercises at a temperature of 1, top-p of 0.95, top-k of 40, and 8192 maximum output tokens.

Only accuracy, calculated as the fraction of correctly answered tasks out of all tasks, has been considered.

After processing the cropped exercises to establish a best-case baseline performance for VQA,

the model selection can be narrowed down. The following subsections discuss the results of these baseline metrics, the aggregated results regarding the impacts of image degradation, and a further discussion.

4.1 Model Selection Using Cropped Exercises

To filter out unviable model candidates, model performance was first assessed under the visual best-case conditions of having visually clear cropped exercises. The results are visible in Table 2. The smaller models performed worse than the expected value for random guessing and are not taken into further consideration.

Also, only two freely available models performed significantly better than the expected value for random guessing.

Their accuracy values were also far behind those of the tested commercial models but are still considered for further evaluation.

GPT-4o and Gemini 1.5 Pro-002 achieved high accuracy ratings of around 87,22% and 84,58% respectively.

Most top-performing models performed better on math problems than on reading comprehension exercises. This is visualized in Figure 2.

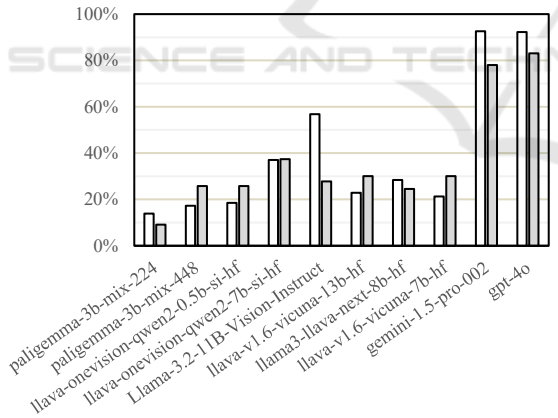

Figure 2: Accuracy on math problems (white) and reading comprehension exercises (grey).

As seen in Figure 3, all models show significant performance differences between answering formats. However, no inter-model trend can be described regarding answer formats.

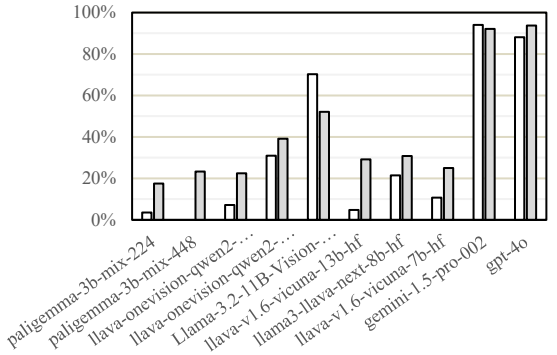

Figure 3: Accuracy on math problems with free response (white) and single choice (grey).

4.2 Impacts of Image Degradation

An aggregated overview of model performances can be seen in Table 3 and is visualized in Figure 4.

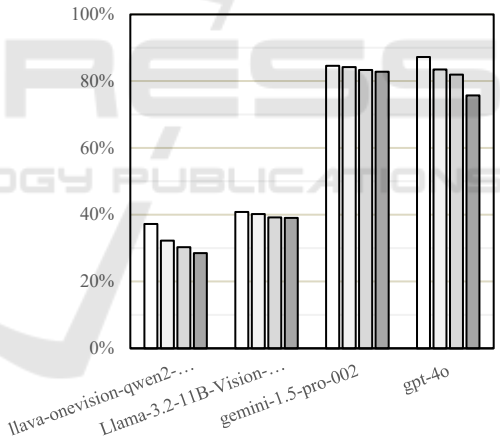

Figure 4: Aggregated performance of selected models on cropped tasks, tasks in sheet context, emulated close-up images, and emulated further-away pictures.

Despite not containing additional exercises, the accuracy of all models decreases when embedding the images in their page context.

Table 3: Accuracy on cropped exercises with selected models highlighted (grey).

Model	Cropped	Sheet	Close-up	Further-away
llava-hf/llava-onevision-qwen2-7b-si-hf	37,22%	32,22%	30,28%	28,47%
meta-llama/Llama-3.2-11B-Vision-Instruct	40,83%	40,20%	39,17%	39,03%
google/gemini-1.5-pro-002	84,58%	84,17%	83,33%	82,78%
openai/gpt-4o	87,22%	83,47%	81,94%	75,69%

Table 4: Accuracy on emulated further-away images with top performing model highlighted (grey).

Model	Math/FRQ	Math/SCQ	Math/Overall	Reading	Overall
llava-hf/llava-onevision-qwen2-7b-si-hf	9,52%	28,75%	23,77%	32,32%	28,47%
meta-llama/Llama-3.2-11B-Vision-Instruct	47,62%	45,42%	45,99%	33,33%	39,03%
google/gemini-1.5-pro-002	86,90%	88,33%	87,96%	78,54%	82,78%
openai/gpt-4o	69,05%	76,67%	74,69%	76,52%	75,69%

This is especially true for the LLaVA-OneVision model and GPT-4o, which have accuracy decreases of 5.00% and 3.75%, respectively.

When a low amount of distortion is introduced to emulate a close-up image, model performances decrease again. However, the effects are more similar for all models this time; accuracy decreases between 0.83% and 1.94%.

Another decrease in performance occurs when emulating further-away images. Again, this is especially true for the LLaVA-OneVision model and GPT-4o, whose accuracy decreases by 6.25% and 1.81%, respectively.

4.3 Discussion

Table 4 shows the final accuracies for emulated further-away images, the scenario modeled closest to the real-life task, aggregated by answer format and exercise area.

Overall, both commercial models maintain a lead in accuracy over the smaller, freely available models.

However, Gemini 1.5 Pro-002, which initially had lower accuracy than GPT-4o on cropped tasks, proves to be more robust in terms of image degradation.

The performance of both commercial models still proves that even under less-than-optimal conditions, both models can successfully be employed for solving exam questions.

This also shows the possibility of students cheating in exams undetected without requiring further expertise or specialized applications.

Furthermore, the study shows the impact of increasing image quality in real-life applications since accuracy scores significantly improve even when just utilizing basic functions of preexisting libraries such as OpenCV.

5 CONCLUSIONS AND PERSPECTIVES

This article aimed to analyse the performance of LVLMs on degraded images of standardized exam questions with three goals:

First, to highlight the feasibility of cheating in in-person exams using smart glasses, to bring awareness to the issue, and to bridge a gap in previous research on using smart devices for this purpose.

In this regard, smart glasses have proven to be a viable input for LVLMs. They also don't require any additional collaboration among students, which rendered previously explored devices unviable. To showcase the feasibility further, a prototype could be developed, for example.

However, the results of the comparative evaluation of the models show a high reliance on commercial cloud models. This does not detract much from the general feasibility of cheating using e.g. smart glasses, since those can already directly interface with cloud models but shows that locally hosted solutions e.g. on smart phones are not yet realistic.

In this context, the feasibility and ease of use is also highlighted through the low performance impacts of image degradation on the tested cloud models. Possibilities for easy-to-achieve performance increases are shown through the use of OpenCV to

Second, the paper was conceived to explore ways of creating benchmarks closer to real-life scenarios and gauging possible performance detriments when moving from synthetic benchmarks to practical applications.

Here, decreases in accuracy have been displayed for various degrees of image degradation compared to ideal conditions. However, more forms and granular degrees of image degradation could be surveyed to quantify the impacts further.

Third, this paper aimed to open the debate on increasing the robustness of exams against new cheating practices utilizing smart devices. While prevention is possible on a surface level by bringing awareness about new developments, so e.g., smart glasses can be spotted and banned more easily, techniques to make inference harder for LVLMs while maintaining the same level of difficulty can be explored to possibly solve the issue on a deeper level.

Exploring exercise formats is a possibility for this. While no inter-model trend could be derived from the data, significant performance differences were shown.

Furthermore, the dataset also includes exercises with figures and tables, which can pose an additional challenge but have not been evaluated in this paper.

When splicing the exams, it stood out that some exercises spanned across multi-columns, leading to wrong outputs when directly exploring OCR capabilities. Exploring possibilities through varying layout options or other supplemental information could thus also prove impactful.

In this regard, the paper also only peripherally addressed the added challenge of having multiple tasks on one page.

Most math problems also named variables directly. To make inference harder for LVLMs, using other terms or synonyms that are obvious to humans could be explored.

Another aspect is the contamination of the training dataset with questions that have been used for benchmarking. This can only be directly evaluated when having access to the training dataset; however, it could be explored whether, e.g., just rephrasing the tasks has an impact, as is known for evaluating decontamination efforts.

While the prevalent goal of exploring these concepts is to increase the robustness of exams, their evaluation could also give further insight into the limits of current models and goals for future models.

REFERENCES

Allam, A., & Haggag, M. (2012). The Question Answering Systems: A Survey. *International Journal of Research and Reviews in Information Sciences*, *2*, 211–221.

Baird, H. S. (2007). The State of the Art of Document Image Degradation Modelling. In S. Singh & B. B. Chaudhuri (Eds.), *Advances in Pattern Recognition. Digital document processing: major directions and recent advances* (pp. 261–279). London: Scholars Portal.

Beyer, L., Steiner, A., Pinto, A. S., Kolesnikov, A., Wang, X. [Xiao], Salz, D., . . . Zhai, X. (2024). *PaliGemma: A versatile 3B VLM for transfer*.

C. Zauner (2010). Implementation and Benchmarking of Perceptual Image Hash Functions.

Chang, Y. [Yupeng], Wang, X. [Xu], Wang, J., Wu, Y. [Yuan], Yang, L. [Linyi], Zhu, K., . . . Xie, X. (2023). *A Survey on Evaluation of Large Language Models*.

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M. [Mark], Jun, H., Kaiser, L., . . . Schulman, J. (2021). *Training Verifiers to Solve Math Word Problems*.

Georgiev, P., Lei, V. I., Burnell, R., Bai, L., Gulati, A., Tanzer, G., . . . Vinyals, O. (2024). *Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context*.

Goyal, Y., Khot, T., Agrawal, A., Summers-Stay, D., Batra, D., & Parikh, D. (2019). Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. *International Journal of Computer Vision*, *127*(4), 398–414.

Guo, Z., Jin, R., Liu, C. [Chuang], Huang, Y., Shi, D., Supryadi, . . . Xiong, D. (2023). *Evaluating Large Language Models: A Comprehensive Survey*.

Guyon, I., Haralick, R. M., Hull, J. J., & Phillips, I. T. (2000). Data sets for OCR and Document Image Understanding Research. In H. Bunke (Ed.), *Handbook of character recognition and document image analysis* (1st ed., pp. 779–799). Singapore: World Scientific.

Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., . . . Steinhardt, J. (2021). *Measuring Mathematical Problem Solving With the MATH Dataset*.

Heya, T. A., Serwadda, A., Griswold-Steiner, I., & Matovu, R. (2021). A Wearables-Driven Attack on Examination Proctoring. In *2021 18th International Conference on Privacy, Security and Trust (PST): 13-15 Dec. 2021* (pp. 1–7). Piscataway, New Jersey: IEEE.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). *Large Language Models are Zero-Shot Reasoners*.

Li, B., Zhang, Y. [Yuanhan], Guo, D. [Dong], Zhang, R., Li, F., Zhang, H. [Hao], . . . Li, C. [Chunyuan] (2024). *LLaVA-OneVision: Easy Visual Task Transfer*.

Liu, Y. [Yuliang], Li, Z., Huang, M., Yang, B., Yu, W., Li, C. [Chunyuan], . . . Bai, X. (2023). *OCRBench: On the Hidden Mystery of OCR in Large Multimodal Models*.

Lu, P., Bansal, H., Xia, T., Liu, J. [Jiacheng], Li, C. [Chunyuan], Hajishirzi, H., . . . Gao, J. (2023). *MathVista: Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts*.

Masry, A., Long, D., Tan, J. Q., Joty, S., & Hoque, E. (2022). ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning. *Findings of the Association for Computational Linguistics: ACL 2022*, 2263–2279.

Mathew, M., Karatzas, D., & Jawahar, C. V. (2020). *DocVQA: A Dataset for VQA on Document Images*.

Meta (2024). Llama 3.2.

Nyamawe, A. S., & Mtonyole, N. (2014). The Use of Mobile Phones in University Exams Cheating: Proposed Solution. *International Journal of Engineering Trends and Technology*, *17*(1), 14–17.

OpenAI (2023). *GPT-4 Technical Report*.

Ray-Ban (2024). Ray-Ban | Meta smart glasses 2024 | Ray-Ban®.

Roy, S., & Roth, D. (2016). *Solving General Arithmetic Word Problems*.

Rupert Urbanski and Ralf Peters (2024). Examining the threat of Smart Glasses to Exam Integrity utilizing LLMs and CV. *AMCIS 2024 TREOs*. (40).

Shen, J., Yin, Y., Li, L. [Lin], Shang, L., Jiang, X., Zhang, M., & Liu, Q. (2021). Generate & Rank: A Multi-task Framework for Math Word Problems. In M.-F. Moens, X. Huang, L. Specia, & S. W. Yih (Eds.), *Findings of the Association for Computational Linguistics: EMNLP*

2021 (pp. 2269–2279). Stroudsburg, PA, USA: Association for Computational Linguistics.

Wong, S., Yang, L. [Lillian], Riecke, B., Cramer, E., & Neustaedter, C. (2017). Assessing the usability of smartwatches for academic cheating during exams. In M. Jones, M. Tscheligi, Y. Rogers, & R. Murray-Smith (Eds.), *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services* (pp. 1–11). New York, NY, USA: ACM.

Wu, Q., Teney, D., Wang, P. [Peng], Shen, C., Dick, A., & van den Hengel, A. (2017). Visual question answering: A survey of methods and datasets. *Computer Vision and Image Understanding*, *163*, 21–40.

Xu, P., Shao, W., Zhang, K. [Kaipeng], Gao, P., Liu, S., Lei, M., . . . Luo, P. (2023). *LVLM-eHub: A Comprehensive Evaluation Benchmark for Large Vision-Language Models*.

Yee, H. (2004). Perceptual Metric for Production Testing. *Journal of Graphics Tools*, *9*(4), 33–40.

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., & Choi, Y. (2019). *HellaSwag: Can a Machine Really Finish Your Sentence?*

Zhou, Y., Guo, C., Wang, X. [Xu], Chang, Y. [Yi], & Wu, Y. [Yuan] (2024). *A Survey on Data Augmentation in Large Model Era*.