
MMMU-PRO NEEDS AN UPDATE

Adam Casson
me@adamcasson.com

March 15, 2026

It is recommended to read this in the original HTML format:
<https://www.adamcasson.com/posts/mmmu-pro-update>

ABSTRACT

MMMU-Pro has become a go-to public eval for frontier labs to showcase general vision capabilities. But the benchmark was created in 2024 and frontier models have improved significantly since then. This paper revisits MMMU-Pro with modern frontier models and finds that: (1) ~43% of questions can be solved by text-only reasoning with current frontier models, (2) models struggle with sheet music transcription (~84% misread rate) despite strong performance on tables and charts, (3) low-resolution pathology images make some questions impossible to answer, and (4) the option augmentation process from 4 to 10 choices introduced additional label noise.

1 Introduction

MMMU-Pro [Yue et al., 2025] has become a go-to public eval for frontier labs to showcase general vision capabilities. For example, it's one of the three multimodal evals Anthropic reports in the Claude Opus 4.6 system card [Anthropic, 2026] and they summarize the dataset pretty concisely:

MMMU-Pro is a multimodal understanding benchmark that tests whether models can correctly perceive, interpret, and reason over college-level questions spanning diverse academic disciplines. MMMU-Pro improved on the original MMMU [Yue et al., 2024] by filtering out text-only solvable questions, expanding multiple-choice options from four to ten, and introducing a vision-only input setting in which questions are embedded directly within images.

	Opus 4.6		GPT-5.2 Thinking		Gemini 3 Pro
Visual reasoning	73.9%	MMMU Pro (no tools)	79.5%	MMMU-Pro	81.0%
MMMU Pro	without tools				
	77.3%	MMMU Pro (w/ Python)	80.4%		
	with tools				

Figure 1: Anthropic, OpenAI, and Google DeepMind all report MMMU-Pro performance as one of their main public multimodal evals.

Frontier models have reported steady improvements in performance on MMMU-Pro. GPT-4o achieved 51.9% accuracy when the paper was published, but today Opus 4.6 gets 73.9%, GPT-5.2 has reached 79.5%, and Gemini 3 Pro hits 81.0%¹.

¹Since the time I ran these evals, Gemini 3 Deep Think was released with 81.5% accuracy. Gemini 3.1 Pro was released which has 80.5% accuracy (which is actually lower than Gemini 3 Pro). And GPT-5.4 Thinking (xhigh) was released with 81.2% accuracy.

But MMMU-Pro was created in 2024, which feels like a decade ago in terms of AI progress. Given its use as a headline multimodal eval by the labs and how much frontier capabilities have advanced in the past 18 months, it’s worth revisiting how well the benchmark holds up and where current models succeed or fail.

2 Text-only frontier models are unreasonably effective

One of the main contributions of the benchmark was filtering out questions that could be solved with text-only. The authors describe this process with the following:

We begin by filtering out questions that can be answered by text-only LLMs. We select four strong open-source LLMs: Llama3-70B-Instruct, Qwen2-72B-Instruct, Yi-1.5-34B-Chat, and Mixtral-8x22B-Instruct and task them with answering the MMMU questions without access to images. The models are required to provide answers even when they indicate that visual input is necessary. We repeat this process ten times for each model, considering a question as “answerable” if a model correctly answers it more than five times. We then exclude any question where at least three out of the four models answer correctly across the majority of trials.

While those were strong open-source models for their day, they are now far behind the knowledge and reasoning capabilities of current models. It’s likely then that there remain text-solvable questions that those models couldn’t solve and thus weren’t filtered out. Even by just looking at Figure 2 from the paper we can see a sample that made it into the final dataset which is completely text-solvable:

Question: A solid copper bar of circular cross section has length $L = 1.25$ m and shear modulus of elasticity $G = 45$ GPa. The bar is designed to carry a $250\text{ N} \cdot \text{m}$ torque acting at the ends. If the allowable shear stress is 30 MPa and the allowable angle of twist between the ends is 2.5° , what is the minimum required diameter?

Original Options:
 (A) 35.7 mm (B) 45.7 mm (C) 55.7 mm

Augmented Options:
 (D) 25.7 mm (E) 40.7 mm (F) 50.7 mm (G) 30.7 mm
 (H) 20.7 mm (I) 60.7 mm

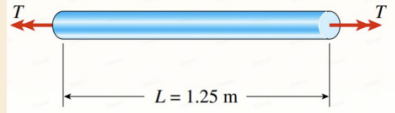


Figure 2: A question presented in the MMMU-Pro paper that is solvable without the image.

Models today though are much more capable than the open-source models of 2024, so to test how the MMMU-Pro text-only filtering stands up to the current frontier models I ran Gemini 3 Pro, GPT-5.2 Thinking, and Claude Opus 4.6 on the Standard (10 options)² version of the benchmark both with and without images³:

Table 1: Model accuracy (%) with and without images on MMMU-Pro Standard (10 options).

Model	w/images	w/o images
Random Choice	12.8	12.8
Gemini 3 Pro	82.5	48.7
GPT-5.2 Thinking	77.8	38.0
Claude Opus 4.6	77.0	46.6

Even without images both Gemini 3 Pro and Opus 4.6 can answer nearly half of MMMU-Pro questions just through textual reasoning alone. I suspect this is possible due to superior reasoning abilities coupled with better world knowledge and recall of these current gen models.

MMMU-Pro does offer a Vision format where the image, question and choices are embedded in a screenshot or photo. I don’t think this fully mitigates the text-only solvable questions, but merely enforces a minimum capability of OCR to

²There are some discrepancies in the accuracies I report here and what the labs report. The standard way MMMU-Pro accuracy is calculated is by averaging model performance on the Standard (10 options) and Vision formats (where question and choices are rendered into a composite image along with the original image). Here I only run the models once on the Standard (10 options) format.

³Gemini 3 Pro evals used gemini-3-pro-preview with thinking level of high. GPT-5.2 Thinking evals used gpt-5.2-2025-12-11 with reasoning effort of high. Claude Opus 4.6 evals used claude-opus-4-6 with thinking type of adaptive and effort of high. The CoT prompt from MMMU-Pro eval was used on all runs.

solve these questions, i.e. if a model can at least read the text in the screenshot then it could ignore the rest of the image content and reason just from its text read-out.

An interesting example of how an unsuspecting question can be solved just with text is in the sample test_Diagnostics_and_Laboratory_Medicine_133. The question given to the model without the image looks like:

Which statement about the figure is correct?

- A. These tumors are usually infrasellar
- B. This tumor represent a metastasis from a lung carcinoma
- C. They always present with endocrine deficiencies
- D. This tumor represents a metastasis from a breast carcinoma
- E. These tumors are usually suprasellar
- F. They never present with endocrine deficiencies

One might think that this can only be answered if the image is given, yet all 3 models solved this question with only text. Opus's output shows how these models can combine their reasoning and knowledge to cleverly deduce the answer just from the options of the question:

Analyzing the Sellar Region Tumor

Step-by-step reasoning:

1. *Key anatomical clue*: The options reference “infrasellar” and “suprasellar,” pointing to a tumor near the sella turcica (pituitary region).
2. *Ruling out metastases*: Options B (lung carcinoma metastasis) and D (breast carcinoma metastasis) can be eliminated—metastases to the sellar region are rare, and the question format suggests a primary tumor.
3. *Endocrine presentation*: Option C (“always present with endocrine deficiencies”) is too absolute—many sellar region tumors present with visual symptoms (bitemporal hemianopia) or headaches rather than endocrine issues. Option F (“never present with endocrine deficiencies”) is also too absolute and incorrect.
4. *Location*: This question most likely depicts a craniopharyngioma, which is: a benign tumor derived from Rathke’s pouch remnants; characteristically suprasellar in location; can compress the optic chiasm and pituitary stalk; may or may not cause endocrine dysfunction.
5. *Infrasellar vs. suprasellar*: Craniopharyngiomas are classically suprasellar, making option A incorrect and option E correct.

Answer: E

Given that so many questions are still text answerable, what happens to these models’ performance if we filter those questions out? We can adopt a similar criterion⁴ as the MMMU-Pro paper by removing questions where at least 2 of the 3 frontier models answer correctly without the image⁵. This results in 43% of questions being removed. We can view the most and least impacted subjects (view full table in Appendix A):

⁴MMMU-Pro’s criteria is stricter as they do 10 runs per model and a question has to be answered correctly a majority of the time by 3/4 models to be filtered out. I wasn’t able to do multiple runs per model because it gets pretty expensive just to run each model once.

⁵I also automatically keep questions that have images as answer options. Upon inspection, the text-only models hallucinate/guess the image contents and get lucky with picking the right option sometimes.

Table 2: Top 5 and bottom 5 subjects by text-only removal rate.

Subject	Total	Removed	Kept	% Removed
Literature	52	39	13	75%
Electronics	60	43	17	72%
Energy and Power	58	39	19	67%
Math	60	37	23	62%
Sociology	54	30	24	56%
Basic Medical Science	52	15	37	29%
Chemistry	60	16	44	27%
Biology	59	12	47	20%
Art	53	9	44	17%
Music	60	9	51	15%
TOTAL	1730	736	994	43%

The subjects with the least amount of image-dependent questions are Literature, Electronics, and Energy and Power while the subjects with the most image-dependent questions are Music, Art, and Biology. This “vision-required” subset is more difficult, the average frontier model performance drops about 10 percentage points compared to the full dataset (view full accuracy delta table in Appendix A):

Table 3: Model accuracy (%) on the full dataset vs. the vision-required subset.

Model	Full (1730)	Vision-Required (994)	Delta
Gemini 3 Pro	82.5	73.8	-8.7
GPT-5.2 Thinking	77.8	68.4	-9.4
Claude Opus 4.6	77.0	65.3	-11.7

Furthermore, we can stratify performance by subject (sorted by average, view full table in Appendix A):

Table 4: Top 5 and bottom 5 subjects by vision-required accuracy (%).

Subject	Gemini 3 Pro	GPT-5.2 Thinking	Claude Opus 4.6
Finance	93.1	96.6	96.6
Electronics	94.1	88.2	88.2
Accounting	86.5	89.2	89.2
Marketing	88.9	85.2	88.9
Public Health	90.2	87.8	80.5
Mechanical Engineering	58.6	51.7	41.4
Agriculture	45.7	42.9	37.1
Literature	46.2	46.2	30.8
Music	43.1	35.3	33.3
Diag. & Lab Medicine	23.7	31.6	21.1
OVERALL	73.8	68.4	65.3

All the models excel at visual tasks like Finance, Electronics, and Accounting, but are all quite bad at subjects like Literature, Music, and Diagnostics and Laboratory Medicine. Notably, those top 5 subjects are similar to each other in distribution of image types but are quite different from the bottom 5 subjects (view full table in Appendix A):

Table 5: Image type distribution for top 5 and bottom 5 subjects by accuracy.

Subject	N	Top Type	2nd Type	3rd Type
Finance	29	Tables (93%)	Plots/Charts (7%)	—
Electronics	17	Diagrams (94%)	Geo. Shapes (6%)	—
Accounting	37	Tables (97%)	Diagrams (3%)	—
Marketing	27	Tables (67%)	Plots/Charts (22%)	Diagrams (11%)
Public Health	41	Tables (83%)	Diagrams (10%)	Plots/Charts (7%)
Mech. Eng.	29	Diagrams (69%)	Blueprints (21%)	Geo. Shapes (7%)
Agriculture	35	Photos (97%)	Microscopic (3%)	—
Literature	13	Photos (46%)	Paintings (38%)	Comics (15%)
Music	51	Sheet Music (100%)	—	—
Diag. & Lab Med.	38	Pathological (53%)	Microscopic (34%)	Medical (21%)

I don’t find this too surprising for a couple reasons. Subjects like finance, accounting, and marketing are all very economically valuable tasks and being good at visual reasoning on images of tables, plots, charts, and diagrams seems like low risk high reward capability to optimize for. It’s probably also easier to optimize for relative to most other image types. High quality tables, plots, charts, and diagrams are plentiful on the web, academic papers, books, etc. They can also be easily generated synthetically, i.e. a strong LLM could create tables, plots, charts, and diagrams using markdown, \LaTeX , HTML, etc. (and create auxiliary data as well, like question-answer pairs about the content) which could then be rendered as images and used in vision encoder and VLM training.

These types of images could also be easier to visually reason about relative to less structured images. For example, on images of tables, the model just needs to have strong OCR skills and 2D spatial understanding, enough to read off the data accurately and reconstruct it in text. Once that’s done the model can basically discard the image and do the rest of its reasoning in text alone.

3 Frontier models can’t read music

What is slightly surprising then is how poorly the models perform on sheet music. It doesn’t seem too far away from the image distribution of tables, plots, charts, and diagrams to explain the drop in performance. After all, if a model had strong enough OCR-like ability to read the musical notation plus 2D spatial understanding it could convert the sheet music to a text-only format and do the rest of its reasoning in text as well. So why are they failing? Do they struggle with reading the music or just lack sufficient music theory knowledge and reasoning?

After looking through the music questions and models’ answers, the common failure mode is that the models fail to read the notes and notation with consistent enough accuracy. When the models attempt to explicitly say what notes or notation they see, they often make obvious mistakes $\sim 84\%$ of the time on average⁶.

Table 6: Sheet music misread rates across frontier models.

Model	Misread	Attempted	Error Rate
Gemini 3 Pro	44	52	85%
GPT-5.2 Thinking	39	46	85%
Claude Opus 4.6	36	44	82%

Take validation_Music_22 for example. The question is:

⁶This was calculated via manual inspection of all 60 Music problems. A misread for a model was noted if it tried to write down what notes/notation it saw in final responses (not thinking tokens/summaries) that included any errors. In some instances, a model could make obvious errors but still get a question correct, i.e. it misreads 2 notes that happen to have the same interval as the notes that the question is about.



Which of the following best describes the seventh chord in the above example?

- A. Major seventh in third inversion
- B. Major/minor seventh in third inversion
- C. Minor/major seventh in first inversion
- D. Minor seventh in third inversion
- E. Major seventh in first inversion
- F. Dominant seventh in second inversion
- G. Dominant seventh in first inversion
- H. Major/minor seventh in second inversion
- I. Minor seventh in second inversion
- J. Diminished seventh in second inversion

All 3 models got this question wrong because they thought they saw just slightly different notes than what was really shown.

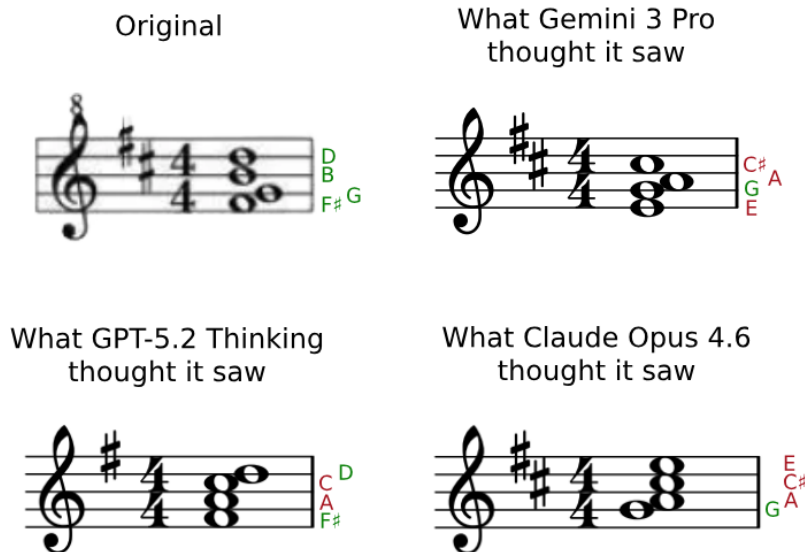


Figure 3: Visualization of how each model transcribed the original sheet music based on their responses. Green annotations indicate notes that each model read correctly, while red are notes they misread.

The original chord is a G major seventh in third inversion.

Gemini and Claude both transcribed the key signature correctly, but both incorrectly transcribed the notes as two variations of the same chord. From bottom note to top Gemini read E-G-A-C# and Claude read G-A-C#-E. Gemini and Claude both reasoned then that the chord must be an A dominant seventh chord (A7) which is also a major minor chord. Gemini then concluded it saw a second inversion of the A7 while Claude concluded it saw a third inversion of the A7. Had the image really shown a chord with notes E-G-A-C# then Gemini's reasoning would've been spot on, and same for Claude if the image really showed G-A-C#-E.

Similar story for GPT-5.2 as well except it misread the key signature, only seeing the F#, and then incorrectly transcribed the chord notes as F#-A-C-D, but the reasoning from there was accurate. It correctly identified that as a D dominant

seventh chord (D7) in first inversion. So once again, had the image really shown F#–A–C–D then GPT-5.2 would’ve reasoned to the right answer.

This pattern of barely misreading some notes (usually off by just one or two staff positions) happens frequently which then sends the model down an unrecoverable route of misguided reasoning. This seems like a lack of finer precision in the 2D spatial understanding of the vision components of the models but without knowing any of the model details it’s hard to tell if it’s an architecture issue or simply a data issue. I imagine stronger real and synthetic data curation could improve this capability a lot, but since reading sheet music isn’t nearly as economically valuable as say finance and accounting tasks, it’s no surprise the models are lacking here.

4 Bad image quality makes some questions impossible

What about the hardest subject, Diagnostics and Laboratory Medicine? The best performing model on the vision-required subset of this subject only hits 31.6% accuracy. Why are these models failing so badly?

From the image type distribution table we can see that Pathological Images and Microscopic Images make up 84% of the images in Diagnostics and Laboratory Medicine which happens to be a modality I’ve worked with for most of my career. Looking at the questions in this subject reveals a lot of low resolution pathology images which is a red flag for the answerability of these questions.

Some brief background, pathology images are digital scans of microscope slides containing tissue samples. This tissue is usually from something like a biopsy or surgical excision that then needs to be examined at the morphological and cellular level by a pathologist in order to render a diagnosis (i.e. determining if a patient has cancer, the type of cancer, the grade of cancer, etc.). In pathology the various scales of the visual diagnostic features can vary greatly, from inspecting the size of cell nuclei to the arrangement of glandular structures, which requires extremely high resolution imagery to capture such features at the same fidelity of various microscope magnification levels.

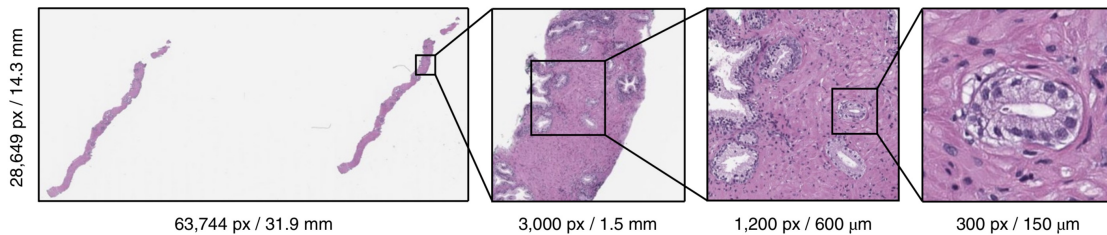
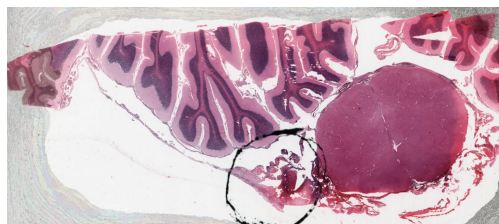


Figure 4: Figure from Campanella et al. [Campanella et al., 2019], original caption: “Left, hematoxylin and eosin slide of a biopsy showing prostatic adenocarcinoma. The diagnosis can be based on very small foci of cancer that account for <1% of the tissue surface. In the slide to the left, only about six small tumor glands are present. The right-most image shows an example of a malignant gland. Its relation to the entire slide is put in perspective to reiterate the difficulty of the task.”

Above is a good illustration from Campanella et al. [Campanella et al., 2019] that shows a full resolution pathology image is of gigapixel scale (possibly even larger with higher resolution scanners nowadays, surpassing $100,000 \times 100,000$ px). However, the presence of cancer could occupy a region less than a 300×300 px. Many of the pathology images in MMMU-Pro try to capture an entire microscope slide in an image resized down to 1900×1600 px and sometimes even as low as 200×150 px in several questions. Clearly, even for larger scale diagnostic features, this resolution is likely to be completely unusable.

A prime example of this is validation_Diagnostics_and_Laboratory_Medicine_20:



45 year old Mexican rancher with 3 month history of cognitive problems. The most likely etiology of this process is:

- A. Coccidioidal meningitis
- B. Toxoplasmosis
- C. Trypanosomiasis
- D. Lyme disease
- E. Tuberculosis meningitis
- F. Amebic encephalitis
- G. Cysticercosis
- H. Cryptococcal meningitis
- I. Herpes Simplex encephalitis
- J. Meningococcal meningitis

The answer is coccidioidal meningitis, but all 3 models predict cysticercosis. To better understand this, fortunately some MMMU-Pro questions have expert explanations of the answers including this one:

This severe basilar meningitis is mediated by coccidioidomycosis. 20–40 micron diameter organisms are identified surrounded by an abundant inflammatory response.

The main takeaway from this explanation is that the discriminatory feature needed to answer this question is the 20–40 micron diameter coccidioides organism. I was able to find a full resolution version of this image via University of Pittsburgh's Department of Neuropathology⁷. Below I zoom into one of the coccidioides a model would need to detect:

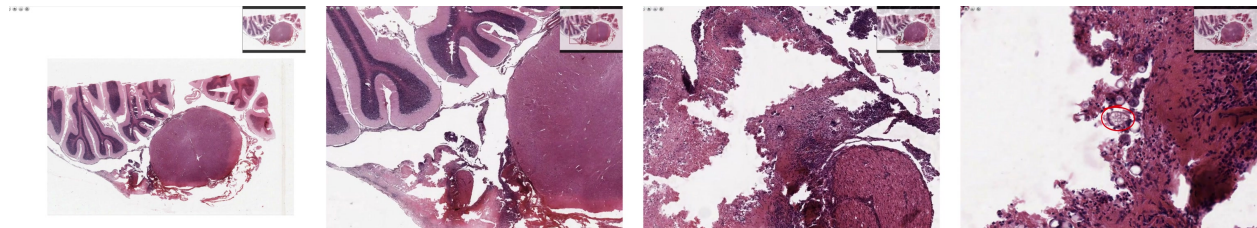


Figure 5: Identifying coccidioides in a pathology image requires high resolution imagery that is able to resolve the organism that is 20–40 microns in size.

In the image given to the model, which is about 1880×840 px, I was able to estimate that the tissue spans about a $45\text{mm} \times 20\text{mm}$ physical area⁸. This would mean one pixel captures a ~ 24 micron square area. These 20–40 micron organisms then are essentially sub-pixel scale in the MMMU-Pro image, which shows this question clearly can't be answered by any model or human!

However, there are some questions in MMMU-Pro that use pathology images that are a zoomed-in crop of a region-of-interest where small diagnostic features become much more visible. Because of this, one hypothesis then is that the models should perform better on these higher magnification ROIs than the low magnification whole slide images.

To test that I manually went through all the questions in Diagnostics and Laboratory Medicine as well as Basic Medical Science and Clinical Medicine to identify all questions that had pathology images⁹ and I categorized them into two buckets of low magnification and high magnification and calculated the models' accuracy:

⁷The case can be seen at <https://neuro2.pathology.pitt.edu/dss/year/1964> and the high resolution viewer of the image at <https://neuro2.pathology.pitt.edu/dss/view/61819b088b53cc0c7af8fbd6/1964-13.svs>. The image is not an exact copy, but is no doubt taken from the same block of tissue and is likely a scan of a slightly different slice of that block.

⁸The full resolution image from UPitt is 62000×39359 px at 0.5 microns per pixel which converts to approximately $31\text{mm} \times 20\text{mm}$. The MMMU-Pro image is 1880×840 px (after cropping padding on the top and bottom). Since the UPitt copy is cropped horizontally compared to the MMMU-Pro version, we can use the vertical dimension to estimate the pixel area of MMMU-Pro image: 20000 microns per 840 px ≈ 24 microns per px. We can then use that to also estimate the physical width of the image: 1880 px $\times 24$ microns per px $\approx 45\text{mm}$.

⁹MMMU-Pro has `img_type` labels for each question but this was bit noisy for this purpose.

Table 7: Model accuracy (%) on pathology images by magnification level.

Model	Low Mag. (N=19)	High Mag. (N=24)
Gemini 3 Pro	5.3	54.2
GPT-5.2 Thinking	15.8	62.5
Claude Opus 4.6	10.5	54.2

And indeed, what we see is the models are performing at around random chance on low magnification pathology images while performing significantly better on the high magnification ROIs.

Since the 19 questions with low magnification images are likely unanswerable due to these images and they make up half of the 38 questions in the vision-required subset of Diagnostics and Laboratory Medicine, this puts an upper bound of 50% accuracy in this subject (not accounting for correct random guessing).

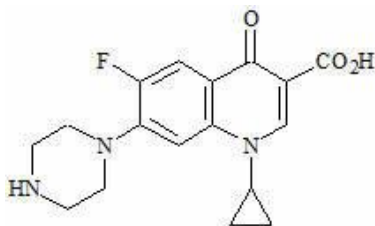
5 Option augmentation increases label noise

Beyond text-only filtering, the other contribution of MMMU-Pro was to expand the number of candidate options of the original MMMU questions from 4 to 10. The authors describe their process for this with the following:

This augmentation is done by human experts with the help of GPT-4o, with additional validation steps to ensure the quality and diversity of the options. Specifically, GPT-4o generates and Claude 3.5 filters the options, followed by two rounds of human review to refine and verify the augmented options.

After sifting through many questions and model responses I noticed some instances of the new candidate options being just as correct as the ground truth answer or even invalidating the ground truth.

One example is validation_Pharmacy_24:



What is the following structure's mechanism of action?

- ◆ A. Alkylating agent
- B. Chain terminator
- ✓ C. Topoisomerase poison
- ◆ D. DNA crosslinker
- ◆ E. RNA polymerase inhibitor
- ◆ F. DNA gyrase inhibitor
- G. Metallating agent
- ◆ H. Reverse transcriptase inhibitor
- ◆ I. DNA intercalator
- J. Antisense agent

Annotations are added to illustrate the ground truth (✓), original distractors in the 4 options format (■) and additional distractors added by MMMU-Pro option augmentation (◆).

All 3 models correctly identify this as the chemical structure of ciprofloxacin and they all select one of the new options, "F. DNA gyrase inhibitor", as their answers. The problem is that this new option happens to also be correct and is arguably a more specific description of ciprofloxacin's mechanism of action. When these new distractors aren't present by evaluating the models on the 4 options format, the models then align with the ground truth by selecting "C. Topoisomerase poison".

To try and quantify noise from option augmentation I flagged questions likely to be noisy based on a simple heuristic where all 3 models agree with each other on the same "wrong" answer. This surfaced 100 questions that I then reviewed

which involved manually solving some questions, consulting online sources (i.e. solutions to textbook questions posted online), verifying the models’ solutions, and when available cross checking the ground truth with explanations provided in the dataset. From this I found that 46 of these seem to have some sort of label noise, $\sim 41\%$ (19) of which could be attributed to noise from option augmentation.

Since the original augmentation process relied on the frontier models available at the time (GPT-4o, Claude 3.5 Sonnet), I suspect current generation models would perform more reliably creating high quality distractors. To test this, I used GPT-5.2 Thinking to re-do option augmentation from scratch for these 19 questions. Given a system prompt, image(s), question, and 4 original options, GPT-5.2 was asked to generate 6 additional options with guidelines to ensure the options were difficult but also not introducing non-mutually exclusive or overly ambiguous distractors (view system prompt in Appendix B).

I then re-evaluated the models on these questions with the newly generated distractors:

Table 8: Model accuracy (%) on noisy subset when reverting from 10 to 4 options and then regenerating 6 more options.

Model	10 options	4 options	Re-augmented 10 options
Gemini 3 Pro	0.0	68.4	63.2
GPT-5.2 Thinking	0.0	73.7	47.4
Claude Opus 4.6	0.0	63.2	57.9

By construction of how this subset of questions was selected, the performance is 0% on the original 10 option format, but we can see that the models recover most of their performance when evaluated on the 4 option format. The re-augmented options using GPT-5.2 appear to mitigate the original 10 option noise but still makes the questions harder than the 4 option format.

These 46 questions represent just a sample of the label noise in the dataset. 27 seem to be pre-existing issues from the original 4 option format and 19 introduced by option augmentation. That’s a 70% increase in noise from the augmentation process in this subset.

Although it is a small sample of just 19 questions with option noise, it may be worth revisiting the augmentation process by using the stronger models that are available today for regenerating the additional candidate options across the full dataset.

6 Conclusion

MMMU-Pro was designed to test the perception, knowledge, and reasoning of multimodal models. But when the image(s) are withheld, current models can answer $\sim 40\%$ of the questions with just knowledge and reasoning alone.

By focusing on the remaining 60% of questions that require perception as well we can get a clearer assessment of model capabilities. The current frontier models show exceptional performance when dealing with tables, plots, and charts, but still struggle in areas like reading sheet music and medical imagery.

To some degree, noise also obscures true model performance. Image noise, like that of the low resolution pathology imagery, can make questions impossible to answer no matter how much models improve. And label noise present in the dataset, some of which is pre-existing and some from option augmentation, also silently penalizes models when they may be correct.

Two of these issues, text-only solvable questions and option augmentation noise, are artifacts of model capabilities in 2024 when the eval was created. The models available at the time that were used for these processes were not strong enough to filter all the text-solvable questions nor be free of error when creating additional options.

MMMU-Pro remains a useful benchmark for multimodal evals, but to maintain relevance it would benefit from re-filtering and re-augmenting options by utilizing the current generation of models.

References

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. In *ACL*, 2025. URL <https://arxiv.org/abs/2409.02813>.

Anthropic. System card: Claude opus 4.6. Technical report, 2026. URL <https://www-cdn.anthropic.com/Odd865075ad3132672ee0ab40b05a53f14cf5288.pdf>.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhua Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*, 2024. URL <https://arxiv.org/abs/2311.16502>.

Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J Busam, Eti Brogi, Victor E Reuter, David S Klimstra, and Thomas J Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, 25(8):1301–1309, 2019. URL <https://www.nature.com/articles/s41591-019-0508-1>.

A Full data tables

Table A1: Full text-only filtering results for all 30 subjects.

Subject	Total	Removed	Kept	% Removed
Literature	52	39	13	75%
Electronics	60	43	17	72%
Energy and Power	58	39	19	67%
Math	60	37	23	62%
Sociology	54	30	24	56%
Marketing	59	32	27	54%
History	56	30	26	54%
Finance	60	31	29	52%
Clinical Medicine	59	30	29	51%
Mechanical Engineering	59	30	29	51%
Physics	60	30	30	50%
Economics	59	27	32	46%
Pharmacy	57	25	32	44%
Art_Theory	55	24	31	44%
Agriculture	60	25	35	42%
Psychology	60	24	36	40%
Architecture and Engineering	60	22	38	37%
Computer Science	60	22	38	37%
Design	60	22	38	37%
Diagnostics and Lab. Medicine	60	22	38	37%
Accounting	58	21	37	36%
Materials	60	21	39	35%
Geography	52	17	35	33%
Manage	50	15	35	30%
Public Health	58	17	41	29%
Basic Medical Science	52	15	37	29%
Chemistry	60	16	44	27%
Biology	59	12	47	20%
Art	53	9	44	17%
Music	60	9	51	15%
OVERALL	1730	736	994	43%

Table A2: Per-subject accuracy delta between full dataset and vision-required subset.

Subject	Gemini 3 Pro	GPT-5.2 Thinking	Claude Opus 4.6	Avg
Literature	-36.5	-36.5	-50.0	-41.0
Diag. & Lab Med.	-26.3	-16.8	-25.6	-22.9
Agriculture	-19.3	-18.8	-19.5	-19.2
Mechanical Engineering	-17.7	-17.8	-21.3	-18.9
Clinical Medicine	-19.6	-16.0	-17.7	-17.8
Psychology	-11.7	-13.9	-20.0	-15.2
Energy and Power	-7.2	-17.8	-17.8	-14.3
Physics	-10.0	-15.0	-16.7	-13.9
Sociology	-11.1	-13.0	-17.6	-13.9
History	-12.6	-8.4	-18.8	-13.3
Math	-8.0	-10.1	-17.1	-11.7
Materials	-6.3	-13.5	-15.3	-11.7
Pharmacy	-12.3	-11.6	-8.8	-10.9
Manage	-9.4	-10.0	-12.0	-10.5
Economics	-6.9	-6.9	-8.6	-7.5
Marketing	-6.0	-8.0	-6.0	-6.7
Art Theory	-6.6	-8.0	-4.8	-6.5
Basic Medical Science	-3.5	-4.0	-9.8	-5.8
Arch. & Eng.	-5.1	-5.1	-5.8	-5.3
Computer Science	-5.1	-2.7	-8.2	-5.3
Geography	-2.7	-2.6	-9.3	-4.9
Chemistry	-3.6	-3.8	-6.7	-4.7
Biology	-1.4	-5.6	-6.9	-4.6
Electronics	-2.5	-6.8	-1.8	-3.7
Music	-5.2	-1.4	-3.3	-3.3
Public Health	-1.1	-1.9	-5.7	-2.9
Design	-1.8	+1.6	+0.4	+0.1
Finance	-0.2	+1.6	-0.1	+0.4
Accounting	+0.3	+1.3	+1.3	+1.0
Art	+0.3	-0.9	+4.8	+1.4
OVERALL	-8.7	-9.4	-11.7	-9.9

Table A3: Full vision-required accuracy (%) for all 30 subjects.

Subject	Gemini 3 Pro	GPT-5.2 Thinking	Claude Opus 4.6
Finance	93.1	96.6	96.6
Electronics	94.1	88.2	88.2
Accounting	86.5	89.2	89.2
Marketing	88.9	85.2	88.9
Public Health	90.2	87.8	80.5
Art	90.9	72.7	84.1
Architecture and Engineering	81.6	81.6	84.2
Economics	81.2	81.2	81.2
Chemistry	86.4	79.5	75.0
Art_Theory	80.6	77.4	80.6
Design	81.6	81.6	73.7
Math	87.0	78.3	69.6
Computer Science	81.6	78.9	68.4
Geography	85.7	68.6	71.4
Energy and Power	84.2	68.4	68.4
Biology	76.6	72.3	66.0
Basic Medical Science	81.1	73.0	59.5
Pharmacy	71.9	65.6	71.9
Physics	76.7	70.0	56.7
Materials	82.1	61.5	56.4
History	73.1	57.7	61.5
Sociology	66.7	66.7	58.3
Manage	68.6	60.0	60.0
Psychology	66.7	52.8	50.0
Clinical Medicine	44.8	58.6	58.6
Mechanical Engineering	58.6	51.7	41.4
Agriculture	45.7	42.9	37.1
Literature	46.2	46.2	30.8
Music	43.1	35.3	33.3
Diagnostics and Lab. Medicine	23.7	31.6	21.1
OVERALL (994)	73.8	68.4	65.3

Table A4: Full image type distribution for all 30 subjects.

Subject	N	Top Type	2nd Type	3rd Type
Finance	29	Tables (93%)	Plots/Charts (7%)	—
Electronics	17	Diagrams (94%)	Geo. Shapes (6%)	—
Accounting	37	Tables (97%)	Diagrams (3%)	—
Marketing	27	Tables (67%)	Plots/Charts (22%)	Diagrams (11%)
Public Health	41	Tables (83%)	Diagrams (10%)	Plots/Charts (7%)
Art	44	Paintings (80%)	Photos (18%)	Portraits (7%)
Arch. & Eng.	38	Tables (58%)	Diagrams (37%)	Geo. Shapes (3%)
Economics	32	Tables (62%)	Plots/Charts (38%)	—
Chemistry	44	Chem. Structures (68%)	Plots/Charts (16%)	Tables (11%)
Art Theory	31	Paintings (48%)	Photos (29%)	Sculpture (19%)
Design	38	Paintings (42%)	Photos (26%)	Diagrams (11%)
Math	23	Diagrams (22%)	Plots/Charts (22%)	Tables (17%)
Computer Science	38	Diagrams (58%)	Trees/Graphs (24%)	Plots/Charts (8%)
Geography	35	Diagrams (34%)	Maps (34%)	Photos (23%)
Energy & Power	19	Diagrams (79%)	Geo. Shapes (16%)	Blueprints (5%)
Biology	47	Diagrams (30%)	Plots/Charts (15%)	Photos (13%)
Basic Med. Sci.	37	Microscopic (24%)	Diagrams (24%)	Medical (19%)
Pharmacy	32	Chem. Structures (56%)	Plots/Charts (16%)	Tables (12%)
Physics	30	Diagrams (77%)	Plots/Charts (17%)	Geo. Shapes (10%)
Materials	39	Diagrams (72%)	Tables (5%)	Geo. Shapes (5%)
History	26	Comics (19%)	Tables (15%)	Maps (15%)
Sociology	24	Photos (29%)	Diagrams (21%)	Comics (17%)
Manage	35	Tables (49%)	Diagrams (17%)	Plots/Charts (14%)
Psychology	36	Plots/Charts (47%)	Diagrams (25%)	Tables (14%)
Clinical Med.	29	Body Scans (52%)	Medical (28%)	Pathological (17%)
Mech. Eng.	29	Diagrams (69%)	Blueprints (21%)	Geo. Shapes (7%)
Agriculture	35	Photos (97%)	Microscopic (3%)	—
Literature	13	Photos (46%)	Paintings (38%)	Comics (15%)
Music	51	Sheet Music (100%)	—	—
Diag. & Lab Med.	38	Pathological (53%)	Microscopic (34%)	Medical (21%)

B Option re-augmentation system prompt

You will be given a question with image(s), its subject area, the existing answer options, and which option is correct. Your job is to generate new distractor options to bring the total to 10.

Rules

1. ****No new option may be correct or arguably correct.**** The ground truth must remain the only right answer.
2. ****No new option should contradict or invalidate the ground truth.**** A test-taker who knows the material should still unambiguously pick the original correct answer.
3. ****All new options must be plausible**** in the context of the question and subject area - they should look like something a student who studied the material might consider.
4. ****College-level difficulty.**** Distractors should not be too easy and still be challenging, including options that are subtly wrong (e.g. a common misconception, an off-by-one value, a related-but-incorrect term).
5. ****Match format, length, and style**** of the existing options. If the originals are short numeric values, produce short numeric values. If they are full sentences, produce full sentences.
6. ****No duplicates**** of existing options or trivial rephrasings

of them.

7. **For quantitative questions**, use numerically plausible values (nearby magnitudes, common calculation errors).
8. **For terminology questions**, use real terms from the same domain that a student might confuse with the correct answer.

Output format

Respond with a single JSON object (no markdown fences):

```
{
  "new_options": ["new option 1", "new option 2", ...],
  "reasoning": "brief explanation of your distractor design
               choices"
}
```

The 'new_options' array must contain exactly the number of new options requested. Do not include letter prefixes (A., B., etc.) - just the raw option text.