DCVLR: Data Curation for Vision Language Reasoning Competition

Oussama Elachqar[*,2] Benjamin Feuer[*,1,2] Neha Hulkund[*,4] Rohun Tripathi[*,3] Rohun Agrawal[†,8] Thao Nguyen[†,5] Nimrod Shabtay[†,15,16] Vishaal Udandarao[†,6,7] Xiaohan Wang[†,1] Stefan Webb[†,2] Yuhui Zhang[†,1] Sara Beery[‡,4] Georgia Gkioxari[‡,8] Paul Liang[‡,4] Ludwig Schmidt[‡,1] Saining Xie[‡,9] Serena Yeung-Levy[‡,1] Michael Buriek[§,12,13] Pavel Bushuyeu[§,12] Dimitrios Damianos[§,10] Eleni Kasoura[§,10,11] Vassilis Katsouros[§,10] Vikas Kumar[§,12] Igor Molybog[§,12] Georgios Paraskevopoulos[§,10] Yosub Shin[§,12] Boris Sobolev[§,12,14] Samuel Watson[§,12] Haoyang Xu[§,12] Emmanouil Koukoumidis (Manos)[α,2]

[*] *Lead Organizer* [α] *Sponsor* [†] *Helping Organizer* [‡] *Advisor* [§] *Participant*

[1] *Stanford University* [2] *Oumi.AI* [3] *AllenAI* [4] *MIT* [5] *Meta* [6] *University of Tübingen* [7] *University of Cambridge* [8] *Caltech* [9] *NYU* [10] *Institute for Language and Speech Processing, Athena Research Center* [11] *Department of Informatics, University of Athens* [12] *University of Hawaii at Manoa* [13] *PwC* [14] *Cisco* [15] *IBM Research* [16] *Tel Aviv University*

DCVLR_NEURIPS@GOOGLEGROUPS.COM

Editors: Tao Qin, Kun Zhang, Jes Frellsen

Abstract

At NeurIPS 2025, we presented the first data-centric, open-data and open-models competition for data curation in vision-language reasoning. The competition focused on improving visual reasoning capabilities of vision-language models (VLMs) through data curation under a fixed supervised-finetuning (SFT) configuration and budget. Submissions were evaluated by fine-tuning a fixed VLM Clark et al. (2026); Bai et al. (2025) on the curated datasets and measuring performance on reasoning benchmarks of multiple-choice visual reasoning questions spanning diverse sources Zhang et al. (2025); Shabtay et al. (2024). The competition provided all necessary resources, including fine-tuning scripts, evaluation code, and baselines. Additionally, as a first of competitions of this kind, organizers also ran the entire SFT and evaluation pipeline on behalf of the contestants, thereby removing a key barrier to entry present in previous data-centric competitions. The result was the largest data-centric competition to date, with 130 signups and 20 submissions. DCVLR centered data curation as the key variable for scientific progress in VLM adaptation, highlighting its importance for real-world applications such as education, biomedicine, and scientific reasoning.

Keywords: data curation, vision-language models, visual reasoning, instruction tuning, data-centric AI, multimodal learning, competition, benchmark

1. Introduction

Visual reasoning has emerged as a central focus in developing next-generation AI Bai et al. (2025); Singh et al. (2025); Marsili et al. (2025); Hudson and Manning (2019); Yue et al. (2024); Lu et al. (2024). A visual reasoning task is any task that requires interpreting and

© 2026 O. Elachqar *et al.*

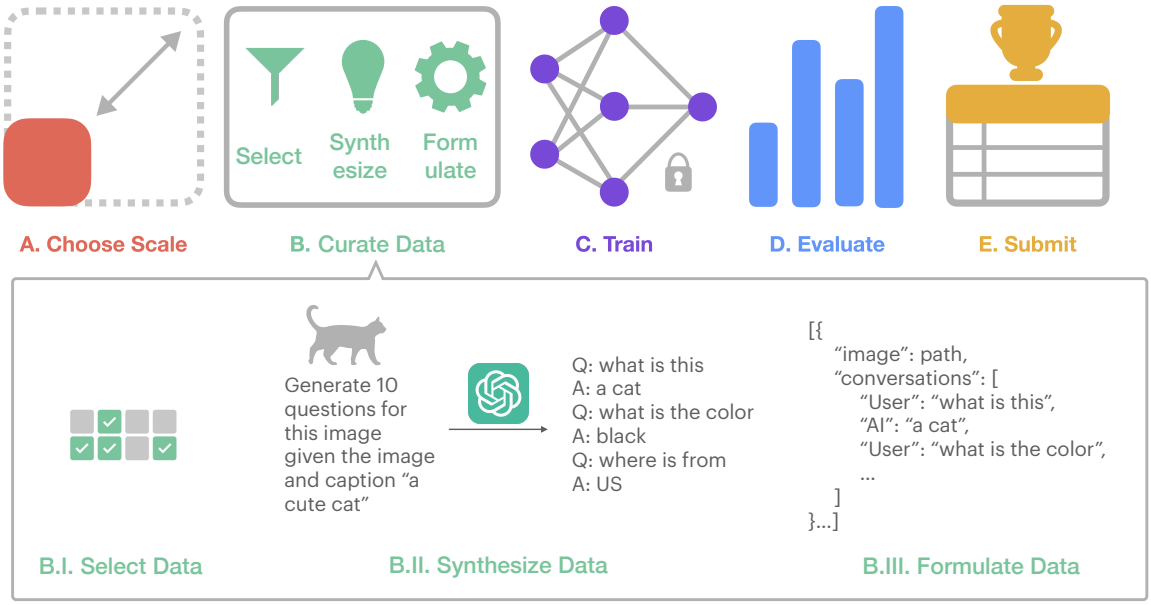


Figure 1: Overview of the DCVLR competition. Participants submitted curated instruction-tuning datasets (up to 10K samples) of their own construction; organizers fine-tuned a fixed, undisclosed vision-language model on each submission and scored it on a suite of held-out visual reasoning benchmarks.

reasoning over visual inputs, and that tends not to saturate with increased model size or pretraining data alone, but instead benefits from improved multi-modal grounding and higher-quality compositional reasoning over image-text interactions. Visual reasoning requires systems to infer causal relationships, interpret spatial configurations, and extract implicit information from visual stimuli. This cognitive leap—from seeing to reasoning—represents a crucial frontier for AI, exemplified in autonomous vehicles that must anticipate a child following a ball, or diagnostic systems that must integrate subtle radiological patterns to form coherent medical hypotheses.

Vision-Language Models (VLMs) like GPT-5 Singh et al. (2025) and Claude-4.7 Anthropic (2026) have shown promising reasoning abilities—for example, interpreting visual scenes or answering complex factoid questions given an image. Yet the question remains: how can we push these capabilities even further?

The Data Curation for Vision Language Reasoning (DCVLR) challenge explored this question through a rigorously controlled data-centric paradigm Gadre et al. (2023); Li et al. (2024). In this challenge, participants were tasked with curating datasets that improved an undisclosed VLM's visual reasoning capabilities. Participants were encouraged to curate data via any strategy that was most effective. The competition imposed only two restrictions: 1) the submitted set had to contain fewer than 10K samples, and 2) no manual annotation was allowed. By fixing the model architecture and training protocol, we isolated and examined the causal relationship between dataset characteristics and reasoning emer-

gence. This approach inverts the traditional model-centric research paradigm, recasting data curation as the primary scientific variable rather than a mere preprocessing step.

Researchers aiming to explore this problem often face significant computational constraints, particularly limited access to adequate compute resources. We removed the computational barriers that typically constrain data curation innovation by providing GPUs to train and evaluate the models. This democratized participation and we attracted a diverse cohort of 130 teams and 20 final submissions spanning students and academic researchers to industry professionals, consistent with participation patterns in analogous data-centric competitions and the rapidly expanding VLM research community. Further to motivate performance and novelty, we established a prize pool of $4,250 USD across three categories.

Fundamentally, this competition reframed data curation as a scientific discipline with its own theoretical questions, methodological approaches, and evaluative criteria. Our competition asks: what data properties catalyze reasoning capabilities in neural systems? By analyzing successful submissions, we extracted principles for effective data curation that transcend specific architectures. These principles offer not merely heuristics but a broader framework for understanding how to systematically instill reasoning abilities in multi-modal systems across domains.

Our best performing curation strategy focused on quality and difficulty of the task rather than quantity, demonstrating previously well-studied quality versus quantity trade-offs Ye et al. (2025); Abdin et al. (2024). The competitors combined two sets to create their high quality 1K training set: a selected set of 900 samples chosen by optimizing for example difficulty and a small set of 100 samples with annotated reasoning traces from an expert model OpenAI (2023). The combined set of 1K samples outperformed other submissions (which used up to 10K samples), demonstrating how a small set of hard high quality samples outperform larger data pools on this task.

2. Related Work

The most comparable existing competitions are DataComp Gadre et al. (2023) and DataComp-LM Li et al. (2024). DataComp focuses on curating data to train better contrastive vision-language models (e.g., CLIP Radford et al. (2021)), while DataComp-LM focuses on language model (LM) training. In both, participants are provided with a massive dataset and are asked to filter and select data to optimize downstream performance.

DCVLR was an entirely new competition centered on data curation for visual reasoning. It offered a data-centric framework for improving the reasoning capabilities of vision-language models (VLMs) by tasking participants with generating high-quality training data, and differed from these prior efforts in three key ways:

1. Focus on data creation, not just selection. Unlike DataComp and DataComp-LM, which restrict participants to selecting from preexisting datasets, we allowed users to bring their own data and only provided references to related datasets that could be used to seed it. Participants were allowed to use any method—retrieval, filtering, augmentation, or synthesis—to generate their data. This design offered far greater creative freedom and encouraged exploration of novel data construction strategies.

2. Targeting reasoning, not general capability. While prior competitions often emphasize general model performance, our competition was specifically designed to probe

visual reasoning. By focusing on this narrower yet critical capability, we aimed to derive actionable insights into how to steer foundation models toward more sophisticated reasoning. These findings can inform downstream applications such as biomedical image understanding, visual tutoring systems, or causal scene interpretation.

3. Fine-tuning over pre-training. Rather than evaluating large-scale pretraining efforts, we focused on fine-tuning existing, high-performing VLMs using curated instruction-tuning data. This significantly reduced computational requirements, making the competition more accessible to academic teams and smaller research groups. It also aligned better with real-world scenarios, where practitioners adapt foundation models to specialized tasks using modest resources.

3. The DCVLR Competition

Rather than focusing on novel architectures or training techniques, DCVLR put reproducible methods for data curation front and center. Participants were tasked with curating up to 10,000 instruction-tuning examples using any strategy and submitting them. The organizers handled all training and evaluation, so participants never needed to run large-scale training themselves.

3.1. Evaluation

The DCVLR evaluation suite consisted of ten benchmarks spanning diverse forms of multimodal reasoning, 10 benchmarks spanning mathematical reasoning, scientific comprehension, and spatial understanding. At the time of the competition, four of the benchmarks were known to participants:

- VMCBench-DEV (Zhang et al., 2025): visual reasoning benchmark that unified 20 existing VQA datasets into a consistent multiple-choice format by simulating errors from different perspectives
- OlympiadBench. (He et al., 2024): bilingual (English and Chinese) competitive Olympiad visual physics and mathematical reasoning questions, each with annotated step-by-step solutions.
- LiveXivVQA (Shabtay et al., 2024): visual question-answer pairs from figures, charts, and tables in recent scientific arXiv papers
- Omni3DBench. (Marsili et al., 2025): non-templated spatial queries on diverse real-world scenes, requiring 3D object grounding and reasoning about distances and dimensions, with four question types: yes/no, multiple choice, counting, and float-point calculations evaluated via mean relative accuracy.

The other six (physical visual understanding) tasks from the Yale Physics Benchmark (Feng et al., 2025) were held-out until final evaluation.

Participants were asked to fine-tune vision-language models from the following options: Qwen2.5-VL-7B (Yang et al., 2024), Molmo (Deitke et al., 2024), or InternVL (Chen et al., 2023). At the time of the competition, the exact evaluation model was deliberately not given

to participants to avoid exploitation of model-specific quirks or formatting practices in data curation. The goal was to ideally find a generalizable visual-reasoning data pool across model families. Final performance was given as the average accuracy over all benchmarks.

3.2. Baselines

We provided 5 key baselines as points of reference for the competition participants:

VL: Visual Learning through Synthetic Generation. This baseline was a multimodal adaptation of the "Simple Scalable Test-Time Compute" paper (Muennighoff et al., 2025), which demonstrated that a small, carefully selected set of reasoning traces could yield strong performance gains. The strategy produced a curated collection of 1,000 high-quality vision-language pairs, generating synthetic data points that targeted common reasoning gaps in existing datasets. We used Claude Sonnet 3.7 to generate Python code that rendered contextually appropriate images (diagrams, geometric figures, charts), which were in turn used to convert text-only reasoning problems into multimodal examples.

LIMO: Language-Informed Multimodal Optimization. This baseline was a VL adaptation of the "Less is More for Reasoning" (LIMO) paper (Ye et al., 2025), which found that models pretrained on broad mathematical knowledge need only a small number of carefully constructed instruction-tuning examples to unlock strong reasoning. The strategy took a language-informed approach to multimodal data synthesis, leveraging Claude Sonnet 3.7 to create contextually rich and semantically coherent vision-language pairs, similarly to the VL baseline. This was also a 1,000-sample synthesis. Both VL and LIMO showed weaker transfer to LiveXivTQA compared to the filtration baselines, suggesting that synthesis alone at small scale does not easily target that distribution.

Multimodal Open R1: Reasoning-focused Filtration. This baseline was derived from the open-source reimplementation of DeepSeek-R1's data generation methodology (DeepSeek-AI et al., 2025), extended to the multimodal setting. It applied reasoning-focused filters to identify data points that specifically enhance multimodal reasoning capabilities, building upon the Open R1 methodology, which emphasizes the importance of reasoning chains and step-by-step problem solving in multimodal contexts. The filtering process prioritized samples that demonstrated clear reasoning patterns and complex visual-textual interactions. The source dataset (`multimodal-open-r1-8192-filtered-mid-ic` on Hugging Face) contains chain-of-thought supervision in R1's characteristic long-form reasoning style, making it potentially strong for tasks requiring explicit step-by-step derivation—though this style also did not generalize as well to the text-heavy LiveXivTQA benchmark.

MM MathInstruct: Mathematical Reasoning at Scale. The MM MathInstruct strategy focused on mathematical reasoning capabilities, filtering from a large corpus of mathematical instruction data (Wang et al., 2025) to identify the most effective samples for enhancing quantitative reasoning in vision-language models. This approach is particularly valuable for applications requiring numerical analysis, geometric reasoning, and mathematical problem-solving across visual and textual domains. It targeted the same OlympiadBench and physics-style reasoning that our competition benchmarks probed, but the results revealed a consistent ceiling: OlympiadBench remained uniformly challenging, with both the baseline Qwen-7B model and fine-tuned models achieving only around 13%

accuracy—suggesting that it requires reasoning capabilities beyond what current data curation strategies can address.

Walton Multimodal Cold Start: Bootstrap Learning Strategy. This strategy addressed the challenging problem of bootstrap learning in multimodal contexts. It was derived from a larger multimodal cold-start corpus (the "Walton" dataset (Oumi AI, 2025)) and filtered by the organizers into a clean 10K set. The Walton strategy achieved a 37.6% relative improvement (from 58.0% to 79.8%) on LiveXivTQA, bringing the fine-tuned model close to Claude Sonnet's performance on that benchmark. The winning team extended this strategy to produce their final result.

For the sake of completeness, we include the model scores from those baseline runs in this proceedings; however, because we, as organizers, inevitably had access to more information about the competition, we ranked our contestants and reported results to them in relation to the **non-fine-tuned base model**, which was Qwen2.5-VL-7B-Instruct.

3.3. Participation

DCVLR ran as an open prize competition from June through October 2025, organized by Oumi.AI and Lambda.AI in collaboration with researchers from Stanford, MIT, NYU, UW, Caltech, AI2, the University of Tübingen, and Tel Aviv University, among others. Registration and evaluation were coordinated through the competition website[1], with technical support provided via a public Discord server and a GitHub starter kit distributed through the Oumi repository.

Engagement spanned a broad funnel from passive interest to full submission. We recorded 130 signups to the competition mailing list and 78 team registrations, of which 20 produced a final dataset submission that was evaluated on our held-out benchmarks. To our knowledge, this makes DCVLR one of the largest data-centric competitions in the vision-language setting to date. Participating teams represented a mixture of academic research groups, industry labs, and independent researchers, consistent with the democratized-access design of the competition: because the organizers handled all fine-tuning and evaluation on shared infrastructure donated by Lambda Labs, contestants only needed modest local resources to prototype and iterate on curation strategies.

A total prize pool of $4,250 USD was awarded across three categories: $3,000 for first place, $1,000 for the Innovation in Data Curation (IDC) prize, and $250 for Honorable Mention. Prize-winning teams were additionally invited to present their methods at the NeurIPS 2025 workshop and to co-author this proceedings. First place was awarded to the University of Hawaii at Manoa (team 3).

Because submissions were open-format BYOD (bring-your-own-data), every submission required automated cleaning before evaluation in order to keep training compute from being wasted on malformed or contaminated data. We processed all submissions through the purpose-built `data-preproc`[2] utility, which normalized input schemas and image sizes, applied text- and image-toxicity filters, and deduplicated submissions against both internal training corpora and our evaluation benchmarks (including the held-out Yale Physics splits). Toxicity and format normalization removed on the order of 1% of samples per submission on

1. https://dcvlr-neurips.github.io/
2. https://github.com/penfever/data-preproc/

average, while benchmark deduplication removed between roughly 5% and 25% of submission content depending on the team's data sources—an empirical confirmation that overlap with public benchmarks is a non-trivial failure mode in data-centric vision-language work, and one that a well-run competition must explicitly control for.

4. Competition Results

4.1. First Place Solution

The University of Hawaii at Manoa, the 1st place finisher of DCVLR, team submitted a compact 1,000-example instruction-tuning dataset constructed by applying difficulty-based filtering to a strongly aligned base corpus. They started from the organizer-provided Walton Multimodal Cold Start dataset, chosen as the base because it exhibits strong distributional overlap with LiveXivTQA — the heavily weighted, text-rich benchmark in the DCVLR evaluation suite — and it is well aligned with the pretrained Qwen2.5-VL-7B-Instruct base model used for fine-tuning. As such, a small data budget could reinforce regions of the task distribution that the model already represents effectively. The primary curation signal was example difficulty, defined as the inconsistency of the base model's answers under stochastic decoding: for each Walton example we ran Qwen2.5-VL-7B-Instruct sixteen times at temperature 0.7 and top-p 0.9, kept examples on which the model was correct on fewer than half of the rollouts, and combined the resulting subset with a small fraction ($\approx 10\%$) of CoSyn-400k examples whose reasoning traces had been rewritten with GPT-4o into longer, more explicit chain-of-thought form. Lightweight token-length filtering and deduplication produced the final 1,000-sample submission.

4.2. Second Place Solution

Team AFIE (Ant Group and HKUST), the 2nd-place finisher in the 1st edition of DCVLR, submitted a four-stage data curation framework titled Multi-Perspective Evaluation and Refinement, aimed at producing high-quality, balanced, and reasoning-intensive multi-modal training data. The team aggregated samples from eight heterogeneous vision-language sources — including LLaVA-OneVision, Multimodal-Open-R1, VLAA-Thinking-SFT, Vision-R1-Cold, MM-Eureka-Distill, R1-Onevision, OpenVLThinker-sft-iter3, OlympiadBench, and BMMR — spanning four task categories (chart QA, document QA, general knowledge QA, and mathematical reasoning QA). In the first stage, they applied stratified balancing jointly over task category and estimated difficulty (easy / medium / hard), intentionally up-sampling hard examples while preserving proportional representation of easier tiers, so the resulting mixture would drive deep reasoning without collapsing into a narrow, over-represented regime.

The second stage introduced a multi-dimensional reject-sampling layer in which four independent LLM/MLLM judges scored every candidate along factual correctness, logical soundness, completeness, and conciseness, with a sample required to clear predefined thresholds on all four axes to pass. Rather than discarding failures, the team routed them into a rejection pool for stage three, where per-sample diagnostic scores and textual feedback from the judges drove a targeted repair loop — automated answer revision to fix factual or logical flaws, and reasoning reconstruction to expand, reorganize, or simplify multi-step

chains — after which refined items were re-submitted to the same four-judge gate and, only if they passed, merged back into the curated pool. The final stage used the combined accepted-plus-refined dataset for SFT and benchmark evaluation.

4.3. Third Place Solution

AthenaRC, the third place finisher of DCVLR, submitted a synthetic multimodal instruction synthesis pipeline, where textual instructions are synthesised based on input images. Starting from the Cauldron collection, they sampled images from the 50 available subsets spanning VQA, OCR, charts, diagrams, spatial reasoning, and mathematical reasoning tasks. They then adapted a Magpie-style self-synthesis framework to the vision-language setting using Qwen2.5-VL-7B-Instruct model. Conditioned on the input image and system prompts, the model first generated user instructions grounded in visual content and subsequently produced paired assistant responses. To improve dataset quality, they applied multiple post-processing stages including sanitization to remove LLM artefacts, formatting validation, duplicate removal, and reward-based filtering. Each generated instruction-response pair was scored using the OpenAssistant DeBERTa-v3 reward model, and only positively scored samples were retained, with the final dataset consisting of the top-ranked 10,000 examples. They found that prompt steering and careful cleaning were especially important, as vision-language models tend to default toward generic captioning behavior without explicit guidance.

5. Results

5.1. Baseline Leaderboard

As shown in Table 1, during our baseline dataset evaluations, performance varied considerably within a fixed model scale, with effective data curation offering a decisive advantage on certain benchmarks, particularly those related to chart and infographic interpretation. This distinction turned out to also be important on our final leaderboard.

5.2. Final Leaderboard

During our baseline dataset evaluations, model capability reliably increased as a function of dataset scale; s1.1-VL and LIMO-vis underperformed compared to larger datasets such as Walton-R1. However, the submission from the first place team clearly demonstrated that scale alone is not a prerequisite for visual reasoning; careful sub-selection from a larger high-quality data pool can produce very strong results in this domain. The large spread of team performance is also noteworthy, considering the relatively small datasets used for the competition, suggesting again that effective data curation can be highly impactful in this regime.

5.3. Why is Walton-R1 a Good Data Source?

Given our findings, it is natural to wonder why the Walton-R1 data performs so well. While we cannot answer this question definitively, we can offer some plausible explanations; compared to the other baseline data sources, which mostly had open-ended responses, Walton-

Baseline	VMC	O3D	OlyB	LXV	YPhys	Avg
MM-Open-R1	77.1	33.3	09.7	66.5	07.1	38.7
s1.1-VL	78.8	33.9	10.2	64.6	05.6	38.6
Molmo-7B-O†	64.6	25.1	01.0	35.5	02.6	25.8
MathInstruct-R1	**80.0**	30.8	07.1	65.8	08.5	38.4
LIMO-vis	79.1	33.5	10.8	65.0	07.9	39.3
Qwen2.5-VL-7B (Base)	79.1	**35.1**	**13.2**	66.6	09.4	40.7
Walton-R1	78.7	34.6	13.1	**79.2**	**10.5**	**43.2**

Table 1: DCVLR baseline comparison across reasoning-oriented fine-tunes of Qwen2.5-VL-7B-Instruct, plus Molmo-7B-O for reference. **Abbreviations:** VMC = VMCBench (dev); O3D = Omni3DBench; OlyB = OlympiadBench; LXV = LiveXiv (unweighted average of the TQA and VQA splits); YPhys = Yale Physics (average of the atomic, electro, mechanics, optics, quantum, and statistics subsets); Avg = macro-average across the five reported benchmarks. †Molmo-7B-O was not evaluated on LiveXiv-TQA, so its LXV entry reflects VQA only. **Findings:** The Walton-R1 SFT mix remained the only curation to improve over the Qwen2.5-VL-7B-Instruct base on the macro-average (43.2% vs. 40.7%, +2.5 absolute); this gap is considerably narrower than a previously reported +10.2 because the earlier spreadsheet reported LiveXiv-TQA-only for most rows but TQA+VQA averages for Walton and MM-Open-R1, artificially inflating those two scores. Under a consistent TQA+VQA average, Walton's advantage comes entirely from LiveXiv (79.2% vs. 66.6%, +12.6), while VMC, Omni3DBench, OlympiadBench, and Yale Physics are essentially tied with the base model. Every other reasoning-SFT variant (LIMO-vis, s1.1-VL, MathInstruct-R1, multimodal-open-R1) under-performed the base on the consistent average, suggesting that naive CoT distillation into a vision-language model can degrade broad VQA capability more than it helps symbolic reasoning. OlympiadBench and Yale Physics remained the hardest benchmarks for every model (all $\leq$14% and $\leq$11% respectively), confirming substantial headroom for future data-curation work on olympiad-level symbolic and physical reasoning.

R1 had a large amount of multiple-choice questions; this may have helped the model on the LiveXiv benchmark, which is multiple choice. Walton-R1 is also the only dataset which contains a significant amount of Chinese (around 11%); since OlympiadBench is dual-lingual, the Chinese content may have helped prevent the regression on that benchmark we observed with other data sources.

6. Conclusion

In this work, we reframed data curation as a scientific discipline with its own theoretical questions, methodological approaches, and evaluative criteria. By fixing the model and training pipeline, the DCVLR challenge isolated the effect of data, enabling a clearer view of how visual reasoning emerges from curated supervision. A key takeaway aligns with findings

Rank	Team	Accuracy (%)	Δ vs Baseline (%)
1	University of Hawaii at Manoa – team 3	46.12	+7.28
2	AFIE	42.13	+3.29
3	AthenaRC	41.81	+2.98
4	ZhuYun – PJLab	38.84	+0.01
5	CoreData	37.39	−1.44
6	Blackwell	37.11	−1.72
7	MICV	34.83	−4.00
8	KDDI Research	20.52	−18.31
–	*Baseline*	*38.83*	-

Table 2: DCVLR Final Leaderboard: NeurIPS 2025 Competition Results. Models were trained and evaluated by the organizers; accuracy is reported on a held-out benchmark and Δ denotes improvement over the provided baseline, Qwen2.5-VL-7B. Qwen2.5-VL-7B scores differ slightly between the original baseline evaluations and the final run because of differences in evaluation dataset composition between the baseline and final leaderboards.

from previous works (Ye et al., 2025; Abdin et al., 2024; Zhou et al., 2023): even small amounts of high-quality, targeted data can significantly improve reasoning performance. Our results suggests that progress in the visual reasoning domain may depend more on identifying the right data than simply scaling quantity.

Looking forward, our proposed paradigm can be extended to new settings. Applying data-centric curation to robotics could shed light on how multimodal experience shapes decision-making, while multilingual extensions would test how reasoning transfers across languages and cultures.

Acknowledgments

The AthenaRC team was funded by the European Union's Horizon Europe research and innovation programme under the AIXPERT project (Grant Agreement No. 101214389). Furthermore, AthenaRC team acknowledges the EuroHPC Joint Undertaking for awarding access to the Leonardo at CINECA, Italy with the EUHPC_D26_056 project. Igor Molybog supported by the National Science Foundation NRT-AI 2244574. Yosub Shin: supported by the National Science Foundation CISE Graduate Fellowships under Grant 2313998. The Hawaii team used cloud GPU resources at NCSA Delta cluster through allocation number CIS240027 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services Support (ACCESS) program, which is supported by National Science Foundation grants 2138259, 2138286, 2138307, 2137603, and 2138296. The technical support and advanced computing resources from University of Hawaii Information Technology Services – Research Cyberinfrastructure, funded in part by the National Science Foundation CC* awards 2201428 and 2232862.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

Anthropic. Introducing claude opus 4.7, 2026. URL https://www.anthropic.com/news/claude-opus-4-7.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Zhong Muyan, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Intern vl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24185–24198, 2023. URL https://api.semanticscholar.org/CorpusID:266521410.

Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Mohammadreza Salehi, Rohun Tripathi, Sangho Lee, Zhongzheng Ren, Chris Dongjoo Kim, Yinuo Yang, Vincent Shao, Yue Yang, Weikai Huang, Ziqi Gao, Taira Anderson, Jianrui Zhang, Jitesh Jain, George Stoica, Winson Han, Ali Farhadi, and Ranjay Krishna. Molmo2: Open weights and data for vision-language models with video understanding and grounding. *arXiv preprint arXiv:2601.10611*, 2026.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Jun-Mei Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiaoling Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bing-Li Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dong-Li Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, JingChang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, M. Tang, Meng Li,

Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, Ruiqi Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shao-Kang Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wen-Xia Yu, Wentao Zhang, Wangding Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyu Jin, Xi-Cheng Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yi Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yu-Jing Zou, Yujia He, Yunfan Xiong, Yu-Wei Luo, Yu mei You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yao Li, Yi Zheng, Yuchen Zhu, Yunxiang Ma, Ying Tang, Yukun Zha, Yuting Yan, Zehui Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhen guo Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zi-Long Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645: 633 – 638, 2025. URL https://api.semanticscholar.org/CorpusID:275789950.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Christopher Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Christopher Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hanna Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 91–104, 2024. URL https://api.semanticscholar.org/CorpusID:274514412.

Kaiyue Feng, Yilun Zhao, Yixin Liu, Tianyu Yang, Chen Zhao, John Sous, and Arman Cohan. Physics: Benchmarking foundation models on university-level physics problem solving. In *Annual Meeting of the Association for Computational Linguistics*, 2025. URL https://api.semanticscholar.org/CorpusID:277434962.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. In *NeurIPS*, 2023.

Chaoqun He, Renjie Luo, Yuzhuo Bai, Shengding Hu, Zhen Leng Thai, Junhao Shen, Jinyi Hu, Xu Han, Yujie Huang, Yuxiang Zhang, Jie Liu, Lei Qi, Zhiyuan Liu, and Maosong Sun. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level

bilingual multimodal scientific problems. In *Annual Meeting of the Association for Computational Linguistics*, 2024. URL https://api.semanticscholar.org/CorpusID:267770504.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Yitzhak Gadre, Hritik Bansal, Etash Guha, Sedrick Scott Keh, Kushal Arora, et al. Datacomp-lm: In search of the next generation of training sets for language models. *Advances in Neural Information Processing Systems*, 37:14200–14282, 2024.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *ICLR*, 2024.

Damiano Marsili, Rohun Agrawal, Yisong Yue, and Georgia Gkioxari. Visual agentic ai for spatial reasoning with a dynamic api, 2025. URL https://arxiv.org/abs/2502.06787.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Fei-Fei Li, Hanna Hajishirzi, Luke S. Zettlemoyer, Percy Liang, Emmanuel J. Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. In *Conference on Empirical Methods in Natural Language Processing*, 2025. URL https://api.semanticscholar.org/CorpusID:276079693.

OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Oumi AI. walton-multimodal-cold-start-r1-format, 2025. URL https://huggingface.co/datasets/oumi-ai/walton-multimodal-cold-start-r1-format. Accessed: 2026-04-28.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. URL https://api.semanticscholar.org/CorpusID:231591445.

Nimrod Shabtay, Felipe Maia Polo, Sivan Doveh, Wei Lin, M. Jehanzeb Mirza, Leshem Chosen, Mikhail Yurochkin, Yuekai Sun, Assaf Arbelle, Leonid Karlinsky, and Raja Giryes. Livexiv – a multi-modal live benchmark based on arxiv papers content, 2024. URL https://arxiv.org/abs/2410.10783.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, Alex Makelov, Alex Neitz, Alex Wei, Alexandra Barr, Alexandre Kirchmeyer, Alexey Ivanov, Alexi Christakis, Alistair Gillespie, Allison Tam, Ally Bennett, Alvin Wan, Alyssa Huang, Amy McDonald Sandjideh, Amy Yang, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrei Gheorghe, Andres Garcia Garcia, Andrew Braunstein, Andrew Liu, Andrew Schmidt, Andrey Mereskin, Andrey

Mishchenko, Andy Applebaum, Andy Rogerson, Ann Rajan, Annie Wei, Anoop Kotha, Anubha Srivastava, Anushree Agrawal, Arun Vijayvergiya, Ashley Tyra, Ashvin Nair, Avi Nayak, Ben Eggers, Bessie Ji, Beth Hoover, Bill Chen, Blair Chen, Boaz Barak, Borys Minaiev, Botao Hao, Bowen Baker, Brad Lightcap, Brandon McKinzie, Brandon Wang, Brendan Quinn, Brian Fioca, Brian Hsu, Brian Yang, Brian Yu, Brian Zhang, Brittany Brenner, Callie Riggins Zetino, Cameron Raymond, Camillo Lugaresi, Carolina Paz, Cary Hudson, Cedric Whitney, Chak Li, Charles Chen, Charlotte Cole, Chelsea Voss, Chen Ding, Chen Shen, Chengdu Huang, Chris Colby, Chris Hallacy, Chris Koch, Chris Lu, Christina Kaplan, Christina Kim, CJ Minott-Henriques, Cliff Frey, Cody Yu, Coley Czarnecki, Colin Reid, Colin Wei, Cory Decareaux, Cristina Scheau, Cyril Zhang, Cyrus Forbes, Da Tang, Dakota Goldberg, Dan Roberts, Dana Palmie, Daniel Kappler, Daniel Levine, Daniel Wright, Dave Leo, David Lin, David Robinson, Declan Grabb, Derek Chen, Derek Lim, Derek Salama, Dibya Bhattacharjee, Dimitris Tsipras, Dinghua Li, Dingli Yu, DJ Strouse, Drew Williams, Dylan Hunn, Ed Bayes, Edwin Arbus, Ekin Akyurek, Elaine Ya Le, Elana Widmann, Eli Yani, Elizabeth Proehl, Enis Sert, Enoch Cheung, Eri Schwartz, Eric Han, Eric Jiang, Eric Mitchell, Eric Sigler, Eric Wallace, Erik Ritter, Erin Kavanaugh, Evan Mays, Evgenii Nikishin, Fangyuan Li, Felipe Petroski Such, Filipe de Avila Belbute Peres, Filippo Raso, Florent Bekerman, Foivos Tsimpourlas, Fotis Chantzis, Francis Song, Francis Zhang, Gaby Raila, Garrett McGrath, Gary Briggs, Gary Yang, Giambattista Parascandolo, Gildas Chabot, Grace Kim, Grace Zhao, Gregory Valiant, Guillaume Leclerc, Hadi Salman, Hanson Wang, Hao Sheng, Haoming Jiang, Haoyu Wang, Haozhun Jin, Harshit Sikchi, Heather Schmidt, Henry Aspegren, Honglin Chen, Huida Qiu, Hunter Lightman, Ian Covert, Ian Kivlichan, Ian Silber, Ian Sohl, Ibrahim Hammoud, Ignasi Clavera, Ikai Lan, Ilge Akkaya, Ilya Kostrikov, Irina Kofman, Isak Etinger, Ishaan Singal, Jackie Hehir, Jacob Huh, Jacqueline Pan, Jake Wilczynski, Jakub Pachocki, James Lee, James Quinn, Jamie Kiros, Janvi Kalra, Jasmyn Samaroo, Jason Wang, Jason Wolfe, Jay Chen, Jay Wang, Jean Harb, Jeffrey Han, Jeffrey Wang, Jennifer Zhao, Jeremy Chen, Jerene Yang, Jerry Tworek, Jesse Chand, Jessica Landon, Jessica Liang, Ji Lin, Jiancheng Liu, Jianfeng Wang, Jie Tang, Jihan Yin, Joanne Jang, Joel Morris, Joey Flynn, Johannes Ferstad, Johannes Heidecke, John Fishbein, John Hallman, Jonah Grant, Jonathan Chien, Jonathan Gordon, Jongsoo Park, Jordan Liss, Jos Kraaijeveld, Joseph Guay, Joseph Mo, Josh Lawson, Josh McGrath, Joshua Vendrow, Joy Jiao, Julian Lee, Julie Steele, Julie Wang, Junhua Mao, Kai Chen, Kai Hayashi, Kai Xiao, Kamyar Salahi, Kan Wu, Karan Sekhri, Karan Sharma, Karan Singhal, Karen Li, Kenny Nguyen, Keren Gu-Lemberg, Kevin King, Kevin Liu, Kevin Stone, Kevin Yu, Kristen Ying, Kristian Georgiev, Kristie Lim, Kushal Tirumala, Kyle Miller, Lama Ahmad, Larry Lv, Laura Clare, Laurance Fauconnet, Lauren Itow, Lauren Yang, Laurentia Romaniuk, Leah Anise, Lee Byron, Leher Pathak, Leon Maksin, Leyan Lo, Leyton Ho, Li Jing, Liang Wu, Liang Xiong, Lien Mamitsuka, Lin Yang, Lindsay McCallum, Lindsey Held, Liz Bourgeois, Logan Engstrom, Lorenz Kuhn, Louis Feuvrier, Lu Zhang, Lucas Switzer, Lukas Kondraciuk, Lukasz Kaiser, Manas Joglekar, Mandeep Singh, Mandip Shah, Manuka Stratta, Marcus Williams, Mark Chen, Mark Sun, Marselus Cayton, Martin Li, Marvin Zhang, Marwan Aljubeh, Matt Nichols, Matthew Haines, Max Schwarzer, Mayank Gupta, Meghan Shah, Melody Huang, Meng Dong, Mengqing Wang, Mia Glaese, Micah Carroll, Michael Lampe, Michael Malek, Michael Sharman,

Michael Zhang, Michele Wang, Michelle Pokrass, Mihai Florian, Mikhail Pavlov, Miles Wang, Ming Chen, Mingxuan Wang, Minnia Feng, Mo Bavarian, Molly Lin, Moose Abdool, Mostafa Rohaninejad, Nacho Soto, Natalie Staudacher, Natan LaFontaine, Nathan Marwell, Nelson Liu, Nick Preston, Nick Turley, Nicklas Ansman, Nicole Blades, Nikil Pancha, Nikita Mikhaylin, Niko Felix, Nikunj Handa, Nishant Rai, Nitish Keskar, Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia Watkins, Oona Gleeson, Pamela Mishkin, Patryk Lesiewicz, Paul Baltescu, Pavel Belov, Peter Zhokhov, Philip Pronin, Phillip Guo, Phoebe Thacker, Qi Liu, Qiming Yuan, Qinghua Liu, Rachel Dias, Rachel Puckett, Rahul Arora, Ravi Teja Mullapudi, Raz Gaon, Reah Miyara, Rennie Song, Rishabh Aggarwal, RJ Marsan, Robel Yemiru, Robert Xiong, Rohan Kshirsagar, Rohan Nuttall, Roman Tsiupa, Ronen Eldan, Rose Wang, Roshan James, Roy Ziv, Rui Shu, Ruslan Nigmatullin, Saachi Jain, Saam Talaie, Sam Altman, Sam Arnesen, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Sarah Yoo, Savannah Heon, Scott Ethersmith, Sean Grove, Sean Taylor, Sebastien Bubeck, Sever Banesiu, Shaokyi Amdo, Shengjia Zhao, Sherwin Wu, Shibani Santurkar, Shiyu Zhao, Shraman Ray Chaudhuri, Shreyas Krishnaswamy, Shuaiqi, Xia, Shuyang Cheng, Shyamal Anadkat, Simón Posada Fishman, Simon Tobin, Siyuan Fu, Somay Jain, Song Mei, Sonya Egoian, Spencer Kim, Spug Golden, SQ Mah, Steph Lin, Stephen Imm, Steve Sharpe, Steve Yadlowsky, Sulman Choudhry, Sungwon Eum, Suvansh Sanjeev, Tabarak Khan, Tal Stramer, Tao Wang, Tao Xin, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Degry, Thomas Shadwell, Tianfu Fu, Tianshi Gao, Timur Garipov, Tina Sriskandarajah, Toki Sherbakov, Tomer Kaftan, Tomo Hiratsuka, Tongzhou Wang, Tony Song, Tony Zhao, Troy Peterson, Val Kharitonov, Victoria Chernova, Vineet Kosaraju, Vishal Kuo, Vitchyr Pong, Vivek Verma, Vlad Petrov, Wanning Jiang, Weixing Zhang, Wenda Zhou, Wenlei Xie, Wenting Zhan, Wes McCabe, Will DePue, Will Ellsworth, Wulfie Bain, Wyatt Thompson, Xiangning Chen, Xiangyu Qi, Xin Xiang, Xinwei Shi, Yann Dubois, Yaodong Yu, Yara Khakbaz, Yifan Wu, Yilei Qian, Yin Tat Lee, Yinbo Chen, Yizhen Zhang, Yizhong Xiong, Yonglong Tian, Young Cha, Yu Bai, Yu Yang, Yuan Yuan, Yuanzhi Li, Yufeng Zhang, Yuguang Yang, Yujia Jin, Yun Jiang, Yunyun Wang, Yushi Wang, Yutian Liu, Zach Stubenvoll, Zehao Dou, Zheng Wu, and Zhigang Wang. Openai gpt-5 system card, 2025. URL https://arxiv.org/abs/2601.03267.

Ke Wang, Junting Pan, Linda Wei, Aojun Zhou, Weikang Shi, Zimu Lu, Han Xiao, Yunqiao Yang, Houxing Ren, Mingjie Zhan, and Hongsheng Li. Mathcoder-vl: Bridging vision and code for enhanced multimodal mathematical reasoning. In *Annual Meeting of the Association for Computational Linguistics*, 2025. URL https://api.semanticscholar.org/CorpusID:278635897.

Qwen An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yi-Chao Zhang, Yunyang Wan, Yuqi Liu, Zeyu Cui, Zhenru Zhang, Zihan Qiu, Shanghaoran Quan, and Zekun Wang. Qwen2.5 tech-

nical report. *ArXiv*, abs/2412.15115, 2024. URL https://api.semanticscholar.org/CorpusID:274859421.

Yixin Ye, Zhen Huang, Yang Xiao, Ethan Chern, Shijie Xia, and Pengfei Liu. Limo: Less is more for reasoning, 2025. URL https://arxiv.org/abs/2502.03387.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*, 2024.

Yuhui Zhang, Yuchang Su, Yiming Liu, Xiaohan Wang, James Burgess, Elaine Sui, Chenyu Wang, Josiah Aklilu, Alejandro Lozano, Anjiang Wei, Ludwig Schmidt, and Serena Yeung-Levy. Automated generation of challenging multiple-choice questions for vision language model evaluation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36:55006–55021, 2023.