NegVQA: Can Vision Language Models Understand Negation?

Anonymous ACL submission

Abstract

Negation is a fundamental linguistic phenomenon that can entirely reverse the meaning of a sentence. As vision language models (VLMs) continue to advance and are deployed in high-stakes applications, assessing their ability to comprehend negation becomes essential. To address this, we introduce *NegVQA*, a visual question answering (VQA) benchmark consisting of 7,379 two-choice questions covering diverse negation scenarios and image-question distributions. We construct *NegVQA* by leveraging large language models to generate negated versions of questions from existing VQA datasets. Evaluating 20 state-of-the-art VLMs across seven model families, we find that these models struggle significantly with negation, exhibiting a substantial performance drop compared to their responses to the original questions. Furthermore, we uncover a U-shaped scaling trend, where increasing model size initially degrades performance on *NegVQA* before leading to improvements. Our benchmark reveals critical gaps in VLMs' negation understanding and offers insights into future VLM development.

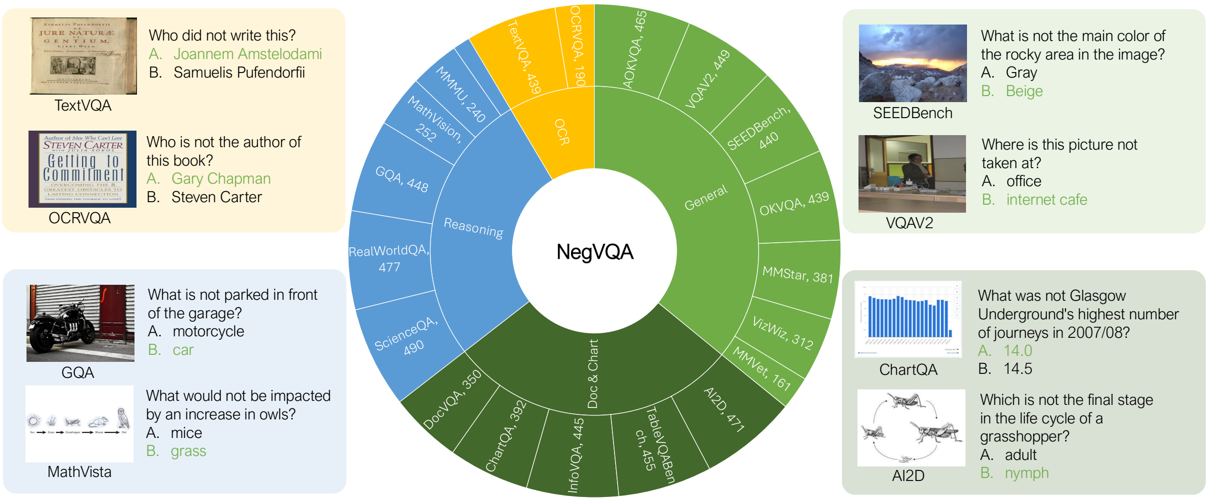
1 Introduction

Vision language models (VLMs) such as GPT-4o and Claude have demonstrated remarkable capabilities in understanding and reasoning about visual content through natural language interactions (OpenAI, 2023; Anthropic, 2024). These models can answer image-based questions, generate descriptions, and engage in multi-turn dialogues about visual scenes (Liu et al., 2023; Deitke et al., 2024; Wang et al., 2024b). More recently, they have been integrated into embodied AI systems and robotics, allowing direct interaction with environments and humans in high-stakes scenarios (Driess et al., 2023; Brohan et al., 2023; Kim et al., 2024a).

Despite their impressive progress, VLMs' ability to understand negation (Ackrill et al., 1975)—a fundamental linguistic phenomenon that can completely alter the meaning of a sentence—remains poorly understood. A failure to correctly interpret negation can lead to critical errors, particularly in interactive AI systems. For instance, if a user instructs a VLM not to take a certain action or asks about something that is absent, misunderstanding negation could result in actions contrary to user intent and pose serious safety risks.

To address this, we introduce *NegVQA*, a visual question answering (VQA) benchmark designed to assess VLMs' comprehension of negation. While existing VQA datasets primarily focus on affirmative questions, *NegVQA* systematically examines negation understanding across diverse scenarios. The dataset consists of 7,379 two-choice questions, covering a range of negation types, including cases where objects are absent, attributes such as colors or sizes are negated, actions are described in terms of what is not happening, and more complex forms of negation that require deeper reasoning. To construct *NegVQA*, we leverage large language models to generate natural negations of questions from existing VQA datasets, ensuring fluency while creating challenging evaluation cases that test both linguistic and visual understanding.

We evaluate 20 state-of-the-art VLMs across seven model families and find that negation remains a major challenge. Despite their strong performance on standard VQA tasks, all models struggle significantly when faced with negated questions. For instance, Qwen2-VL-72B (Wang et al., 2024b), the best-performing model, achieves 92.2% accuracy on original questions but drops nearly 20 percentage points to 72.7% on *NegVQA*. Furthermore, we observe a U-shaped scaling trend, where increasing model size initially leads to worse performance on negation before eventually improving. This finding raises important questions about how VLMs process negation and how to scale up VLMs to enhance negation understanding abilities.



Who did not write this?
A. Joannem Amstelodami
B. Samuelis Pufendorfii

Who is not the author of this book?
A. Gary Chapman
B. Steven Carter

What is not parked in front of the garage?
A. motorcycle
B. car

What would not be impacted by an increase in owls?
A. mice
B. grass

What is not the main color of the rocky area in the image?
A. Gray
B. Beige

Where is this picture not taken at?
A. office
B. internet cafe

What was not Glasgow Underground's highest number of journeys in 2007/08?
A. 14.0
B. 14.5

Which is not the final stage in the life cycle of a grasshopper?
A. adult
B. nymph

Figure 1: ***NegVQA* dataset overview.** *(Middle) NegVQA* comprises a diverse set of negated questions, totaling 7,379 instances sourced from various VQA datasets and domains (general, document/chart, reasoning, OCR). *(Left/Right)* Example questions from different datasets and domains, with correct answers highlighted in green.

In summary, we propose *NegVQA*, a critical diagnostic tool for evaluating negation comprehension in VLMs. Our study establishes baseline performance across major VLM families, reveals their significant shortcomings and uncovers scaling behaviors. These insights highlight the need to develop more robust and trustworthy VLMs that can accurately handle negation, a fundamental aspect of natural language understanding.

2 Dataset: *NegVQA*

This section details the construction and statistics of *NegVQA*, our benchmark for evaluating vision language models' ability to handle negation.

2.1 Data Curation

We construct *NegVQA* by systematically transforming questions from VMCBench (Zhang et al., 2025), a multi-choice visual question answering (VQA) benchmark spanning various datasets and domains, into negated versions using GPT-4o (OpenAI, 2023). Our curation process consists of two main steps.

First, we prompt GPT-4o to generate negated versions of the original questions while preserving their syntactic structure and meaning (see Appendix Figure 3 for prompt details). For example, the question *"Who wrote this book?"* is transformed into *"Who did not write this book?"* We exclude questions that cannot be meaningfully negated (e.g., *"Find the value of x."*), as determined by GPT-4o's assessment of their negatability. After filtering, 7,379 out of 9,018 questions were identified as negatable and successfully transformed. To assess the accuracy of GPT-4o's negation process, we manually verified 100 sampled negated questions and found that 97% were correctly negated (three errors are provided in Appendix Figure 4), confirming the high reliability of the method.

Second, we adjust the answer choices to reflect the negation. Each original four-choice question is reduced to a two-choice format, where we select the correct answer and randomly sample an incorrect choice, then invert their correctness. For instance, in the original question *"Who wrote this book?"*, if the correct answer is *"Samuelis Pufendorfii"* and an incorrect choice is *"Joannem Amstelodami"*, we generate *"Who did not write this book?"* where *"Joannem Amstelodami"* becomes the correct answer, and *"Samuelis Pufendorfii"* the incorrect one. This ensures that the negation meaningfully impacts the answer selection.

2.2 Statistics and Examples

NegVQA incorporates questions from 20 widely-used VQA datasets within VMCBench, covering a broad range of vision language understanding tasks. It includes datasets for **general VQA capabilities** (VQAv2 (Goyal et al., 2017), OKVQA (Marino et al., 2019), MMVet (Yu et al., 2024), VizWiz (Gurari et al., 2018), A-OKVQA (Schwenk et al., 2022), MMStar (Chen et al., 2024), SEEDBench (Li et al., 2024)), **reasoning tasks** (MathVision (Wang et al., 2024a), GQA (Hudson and Manning, 2019), MMMU (Yue et al., 2024), RealWorldQA (xAI, 2024), MathVista (Lu et al., 2024b), ScienceQA (Lu et al., 2022)), **OCR-based VQA**

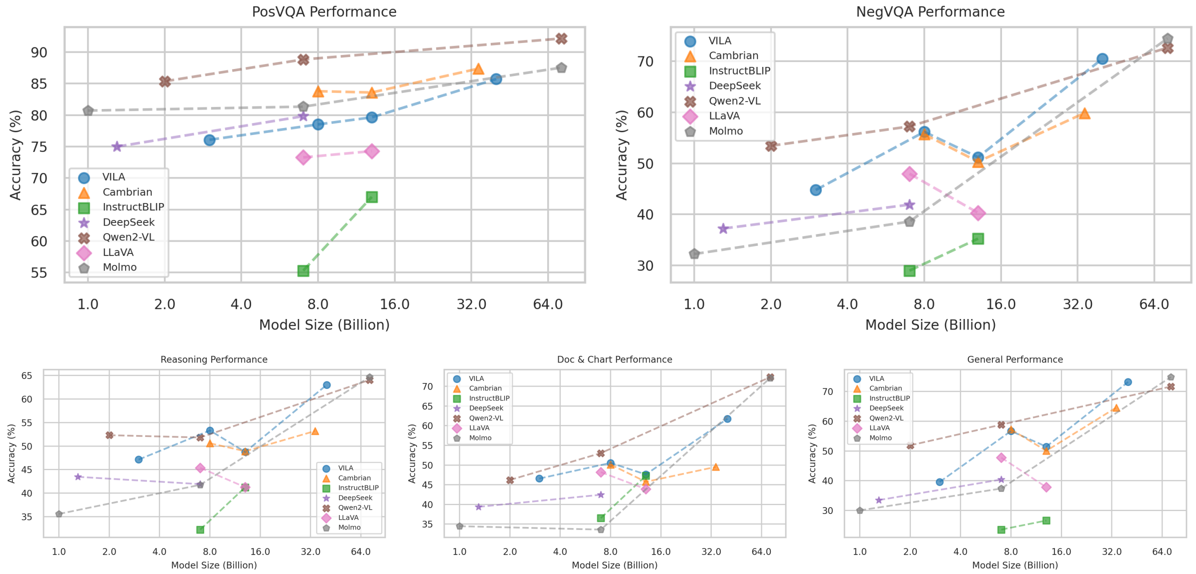


Figure 2: **Model performance and scaling analysis on *NegVQA* across different VLM families and task categories.** *(Top left)* Performance on the original non-negated two-choice questions shows high accuracy and a positive scaling trend. *(Top right)* Performance on *NegVQA* (negated two-choice questions) is significantly lower, with models exhibiting a U-shaped scaling pattern—initially decreasing before improving as model size increases. *(Bottom)* Category-wise breakdown of *NegVQA* performance (reasoning, document/chart, general), where the U-shaped scaling effect is more pronounced in reasoning and document/chart categories.

(OCRVQA (Mishra et al., 2019), TextVQA (Singh et al., 2019)), and **document/chart comprehension** (DocVQA (Mathew et al., 2021), InfoVQA (Mathew et al., 2022), ChartQA (Masry et al., 2022), TableVQABench (Kim et al., 2024b), AI2D (Kembhavi et al., 2016)). The final dataset contains 7,379 questions distributed across these datasets and domains, with the detailed distribution and example questions visualized in Figure 1.

NegVQA is designed to systematically test VLMs' ability to process negation in diverse visual scenarios. The dataset ensures diversity in negation forms, covering cases related to objects, attributes, logical reasoning, spatial relationships, and more. Additionally, all transformed questions have strong visual relevance, requiring models to understand both the image content and the linguistic negation to generate correct answers. *NegVQA* thus serves as a comprehensive benchmark that evaluates vision language models' ability to understand negation in different visual scenarios, providing critical insights into their limitations and potential improvements.

3 Results

In this section, we describe our experimental setup and present our findings on VLM performance on *NegVQA*. Our evaluation highlights two key insights: current VLMs exhibit significant difficulty in understanding negation, regardless of their size or architecture, and model scaling exhibits a U-shaped performance trend.

3.1 Experimental Setup

We evaluated 20 state-of-the-art vision language models (VLMs) from 7 model families on *NegVQA*, including Qwen2-VL (Wang et al., 2024b), Molmo (Deitke et al., 2024), Cambrian (Tong et al., 2024), VILA (Lin et al., 2024), DeepSeek-VL (Lu et al., 2024a), LLaVA1.5 (Liu et al., 2023), and InstructBLIP (Dai et al., 2023). For each family, we tested multiple model sizes to analyze scaling behavior. All evaluations were conducted in a zero-shot setting using the prompt:

```
Question: <image> {question}
Options: A. {A} B. {B} C. {C} D. {D}
Answer with the option's letter from the
given choices directly.
```

The results are summarized in Figure 2, with detailed performance provided in Appendix Table 1.

3.2 Findings

VLMs struggle with negation understanding. Our evaluation reveals that current VLMs consistently underperform on *NegVQA* compared to standard, non-negated VQA tasks. As shown in Figure 2 (top left vs. top right), performance drops significantly across all model families on negated questions. The highest-performing model, Qwen2-VL-72B (Wang et al., 2024b), achieves only 72.7% accuracy on *NegVQA*, compared to 92.2% on non-negated questions—a gap of 19.5 percentage points. On average, model performance decreases by 29.7 points on negated questions compared to the original non-negated questions. This substantial decline is observed across different question types and domains, indicating a fundamental limitation in how VLMs process negation. Appendix Table 1 provides detailed numerical results.

Model scaling exhibits a U-shaped trend. An intriguing pattern emerges in model scaling: as models grow larger, their performance on *NegVQA* initially degrades before improving at the highest scales. This U-shaped trend (Wei et al., 2022; Zhang et al., 2023) is demonstrated in model families such as Cambrian (Tong et al., 2024) and VILA (Lin et al., 2024) (Figure 2, top right), and is more evident in reasoning and document/chart-based tasks (Figure 2, bottom left). Appendix Figure 5 provides a detailed breakdown of performance across individual datasets.

The U-shaped scaling behavior can be interpreted into three phases. In the initial phase, smaller models exhibit limited but relatively stable performance on *NegVQA*. In the intermediate phase, as models scale up, their accuracy declines—likely because they become more proficient at answering standard VQA questions but fail to adjust for negation, leading them to misinterpret negated queries as affirmative ones. Finally, in the large-scale phase, models begin to recover, demonstrating improved negation comprehension, likely due to the development of more advanced language understanding capabilities.

Overall, these results underscore the persistent challenges VLMs face in handling negation and highlight the intriguing scaling behavior of VLMs.

4 Related Work

Vision language models (VLMs). VLMs enable multimodal understanding by modeling $p(y_t|y_{<t}, x)$ in an auto-regressive manner, where y_i represents text tokens and x represents visual input. Modern VLMs typically comprise three key components: a visual encoder (often CLIP (Radford et al., 2021)), a language model, and a linear or MLP projector connecting them. Notable examples include proprietary models such as GPT-4o (OpenAI, 2023) and Claude (Anthropic, 2024), as well as open-source models like LLaVA (Liu et al., 2023) and BLIP (Li et al., 2023). These models are generally trained on image-text pairs and instruction-tuning datasets, leveraging pre-trained vision and language components. While they exhibit strong performance on various image understanding tasks (Liu et al., 2023; Deitke et al., 2024; Wang et al., 2024b) and have been applied in embodied AI and robotics (Driess et al., 2023; Brohan et al., 2023; Kim et al., 2024a), their ability to handle negation remains largely unexplored.

Negation understanding. Negation plays a fundamental role in language comprehension (Ackrill et al., 1975). Most prior research has focused on evaluating language models' ability to understand negation (Hossain et al., 2020; Fancellu and Webber, 2015; Kassner and Schütze, 2020; Zhang et al., 2023). More recently, studies have begun assessing CLIP (Radford et al., 2021)'s understanding of negation (Alhamoud et al., 2025; Singh et al., 2024; Quantmeyer et al., 2024). However, to the best of our knowledge, no prior work has systematically evaluated negation comprehension in generative VLMs. In this work, we introduce *NegVQA*, the first benchmark designed to assess VLMs' ability to handle negation. Given the increasing deployment of VLMs in real-world embodied AI systems, understanding their limitations in processing negation is crucial, as failures in user intent interpretation could lead to unintended and risky scenarios.

Scaling trends. Scaling up models has been a dominant approach in advancing foundation models. However, most scaling studies have focused on language models (Kaplan et al., 2020; Brown et al., 2020; Ruan et al., 2024). While many tasks benefit from scaling, some exhibit inverse scaling (Lin et al., 2022; McKenzie et al., 2023) or U-shaped scaling (Wei et al., 2022; Zhang et al., 2023). In this work, we analyze scaling effects in vision language models on the negation task and reveal a similar U-shaped scaling pattern.

5 Conclusion

In this work, we present *NegVQA*, a benchmark designed to evaluate vision language models' ability to comprehend negation. Our analysis of 20 VLMs highlights their significant limitations in handling negation and uncovers a U-shaped scaling pattern in performance. We envision *NegVQA* as a valuable resource for advancing linguistically competent, safe, and trustworthy vision language models.

Limitations

Our study has three limitations: First, while our multiple-choice format enables controlled experimentation and easy evaluation metrics, it may not fully capture how VLMs handle negation in more open-ended or real-world scenarios where models cannot rely on predefined answer choices. Second, we focus exclusively on zero-shot evaluation, due to current VLMs' architectural constraint of accepting only single image inputs, leaving unexplored how few-shot prompting might affect negation understanding and performance scaling. Third, although we manually verified the accuracy of 97% of our automatically generated questions, our LLM-based approach for converting existing VQA questions into negated forms may introduce subtle errors in question formulation. Despite these limitations, our work provides the first comprehensive analysis of how VLMs process negation, uncovering both their current limitations and a U-shaped scaling pattern. The *NegVQA* benchmark establishes a foundation for systematically evaluating and improving how future vision language models handle this fundamental linguistic operation.

References

John L Ackrill et al. 1975. *Categories and De interpretatione*. Clarendon Press.

Kumail Alhamoud, Shaden Alshammari, Yonglong Tian, Guohao Li, Philip Torr, Yoon Kim, and Marzyeh Ghassemi. 2025. Vision-language models do not understand negation. *arXiv preprint arXiv:2501.09425*.

Anthropic. 2024. Introducing the next generation of claude.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, et al. 2023. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *arXiv preprint arXiv:2307.15818*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *NeurIPS*.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. 2024. Are we on the right way for evaluating large vision-language models? In *NeurIPS*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. 2024. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. 2023. Palm-e: An embodied multimodal language model. In *ICML*.

Federico Fancellu and Bonnie Webber. 2015. Translating negation: A manual error analysis. In *Proceedings of the Second Workshop on Extra-Propositional Aspects of Meaning in Computational Semantics (ExProM 2015)*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *CVPR*.

Md Mosharaf Hossain, Venelin Kovatchev, Pranoy Dutta, Tiffany Kao, Elizabeth Wei, and Eduardo Blanco. 2020. An analysis of natural language inference benchmarks through the lens of negation. In *EMNLP*.

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

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Nora Kassner and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *ACL*.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *ECCV*.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. 2024a. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*.

Yoonsik Kim, Moonbin Yim, and Ka Yeon Song. 2024b. Tablevqa-bench: A visual question answering benchmark on multiple table domains. *arXiv preprint arXiv:2404.19205*.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. 2024. Seed-bench: Benchmarking multimodal large language models. In *CVPR*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. 2024. Vila: On pre-training for visual language models. In *CVPR*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring how models mimic human falsehoods. In *ACL*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *NeurIPS*.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. 2024a. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*.

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

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *NeurIPS*.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *CVPR*.

Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *ACL Findings*.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. 2022. Infographicvqa. In *WACV*.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *WACV*.

Ian R McKenzie, Alexander Lyzhov, Michael Martin Pieler, Alicia Parrish, Aaron Mueller, Ameya Prabhu, Euan McLean, Xudong Shen, Joe Cavanagh, Andrew George Gritsevskiy, et al. 2023. Inverse scaling: When bigger isn't better. *TMLR*.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. Ocr-vqa: Visual question answering by reading text in images. In *ICDAR*.

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

Vincent Quantmeyer, Pablo Mosteiro, and Albert Gatt. 2024. How and where does clip process negation? *arXiv preprint arXiv:2407.10488*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Yangjun Ruan, Chris J Maddison, and Tatsunori Hashimoto. 2024. Observational scaling laws and the predictability of language model performance. *arXiv preprint arXiv:2405.10938*.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *ECCV*.

Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *CVPR*.

Jaisidh Singh, Ishaan Shrivastava, Mayank Vatsa, Richa Singh, and Aparna Bharati. 2024. Learn" no" to say" yes" better: Improving vision-language models via negations. *arXiv preprint arXiv:2403.20312*.

Shengbang Tong, Ellis L Brown II, Penghao Wu, Sanghyun Woo, ADITHYA JAIRAM IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. 2024. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. In *NeurIPS*.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. 2024a. Measuring multimodal mathematical reasoning with math-vision dataset. *arXiv preprint arXiv:2402.14804*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Jason Wei, Yi Tay, and Quoc V Le. 2022. Inverse scaling can become u-shaped. *arXiv preprint arXiv:2211.02011*.

xAI. 2024. Realworldqa dataset.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2024. Mm-vet: Evaluating large multimodal models for integrated capabilities. In *ICML*.

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

Yuhui Zhang, Yuchang Su, Yiming Liu, Xiaohan Wang, James Burgess, Elaine Sui, Chenyu Wang, Josiah Aklilu, Alejandro Lozano, Anjiang Wei, et al. 2025. Automated generation of challenging multiple-choice questions for vision language model evaluation. *arXiv preprint arXiv:2501.03225*.

Yuhui Zhang, Michihiro Yasunaga, Zhengping Zhou, Jeff Z HaoChen, James Zou, Percy Liang, and Serena Yeung. 2023. Beyond positive scaling: How negation impacts scaling trends of language models. In *ACL 2023*.

	Original Non-negated Questions					Negated Questions (*NegVQA*)				
Model	**General**	**Reason**	**OCR**	**Doc&Cht**	**Average**	**General**	**Reason**	**OCR**	**Doc&Cht**	**Average**
Cambrian-8B	87.6	74.0	93.4	80.9	83.8	57.2	50.6	71.8	50.1	55.7
Cambrian-13B	87.7	73.5	95.9	80.5	83.6	50.1	48.9	69.2	45.7	50.3
Cambrian-34B	90.0	80.3	96.6	85.0	87.4	64.6	53.2	81.5	49.5	59.9
InstructBLIP-7B	58.5	53.9	70.0	48.4	55.3	23.6	32.2	20.7	36.5	28.9
InstructBLIP-13B	75.8	62.5	68.1	53.9	67.0	26.6	41.2	20.3	47.3	35.2
DeepSeek-VL-1.3B	81.5	66.6	88.6	65.9	75.0	33.5	43.5	34.6	39.4	37.2
DeepSeek-VL-7B	84.7	71.2	91.3	73.1	79.8	40.4	41.9	53.7	42.4	41.9
LLaVA-1.5-7B	81.0	67.7	85.5	61.1	73.3	47.7	45.4	49.7	48.2	47.9
LLaVA-1.5-13B	82.8	66.5	86.4	62.3	74.3	37.8	41.2	40.4	43.9	40.3
Molmo-1B	83.6	71.7	92.0	77.7	80.7	30.0	35.6	30.4	34.5	32.2
Molmo-7B-O	83.1	69.9	91.2	81.4	81.3	37.4	41.7	49.4	33.6	38.6
Molmo-7B-D	85.6	67.8	94.8	84.3	83.0	55.9	48.6	75.3	49.7	55.3
Molmo-72B	89.4	78.2	96.7	89.0	87.5	74.8	64.7	93.9	72.1	74.5
Qwen2-VL-2B	88.6	74.7	96.1	84.8	85.4	51.9	52.3	78.0	46.2	53.4
Qwen2-VL-7B	91.3	79.8	97.2	89.4	88.8	58.8	51.8	82.0	53.0	57.2
Qwen2-VL-72B	93.6	83.4	99.0	94.8	92.2	71.7	64.1	91.8	72.4	72.7
VILA1.5-3B	83.9	68.0	88.2	66.0	76.1	39.6	47.1	51.9	46.6	44.8
VILA1.5-8B	85.3	71.2	91.0	69.4	78.5	56.7	53.3	68.4	50.5	56.2
VILA1.5-13B	85.7	73.7	91.6	70.3	79.6	51.4	48.7	62.5	47.6	51.2
VILA1.5-40B	89.4	78.6	96.3	81.5	85.7	73.2	63.0	90.3	61.8	70.5

Table 1: **Performance of 20 vision language models from 7 families on NegVQA and the original non-negated dataset.**

```
**Task:**
You will be given an question collected from existing visual question answering datasets. Your task is to
     ↪ produce a minimally modified, negated version of the question by inserting a negation (e.g., "not",
     ↪ "do not", "isn't", etc.) in a way that:

1. **Minimal Changes:** Alters the original question as little as possible.
2. **Answer Inversion:** Causes the original correct answer to become incorrect while making one of the
     ↪ originally incorrect answers correct.
3. **Linguistic Accuracy:** Adheres to proper grammar and preserves the semantic intent of the question.

**Special Case:**
1. Do not negate any background that is provided along with the question (e.g., mathematical conditions,
     ↪ background information, etc). Only negate the question itself (usually the last sentence).
2. If it is not possible to create a valid negation that meets these criteria, return an empty string for
     ↪ the negated question and set the flag `is_negatable` to `false`.

**Output Format:**
Your response should be an object with the following structure:
{
  "negated_question": "<your negated question (with original background information) here, or an empty
       ↪ string if not negatable>",
  "is_negatable": <true/false>
}
```

Figure 3: **Detailed prompts for adding the negation using GPT-4o.**

```
Original Question: how many total singles does he have?
Negated Question: how many total singles does he not have?

Original Question: As shown in the figure, points A, B, and C are three points on O, and the straight line
     ↪ CD and O are tangent to point C. If DCB = 40.0, then the degree of CAB is ()
Negated Question: As shown in the figure, points A, B, and C are three points on O, and the straight line
     ↪ CD and O are not tangent to point C. If DCB = 40.0, then the degree of CAB is ()

Original Question: If cricket was removed from the food web, there would be
Negated Question: If cricket was not removed from the food web, there would be
```

Figure 4: **Errors in negated questions generated by GPT-4o.** The first question cannot be negated, while the second and third questions are negated in the condition, whereas the negation should apply to the main question.

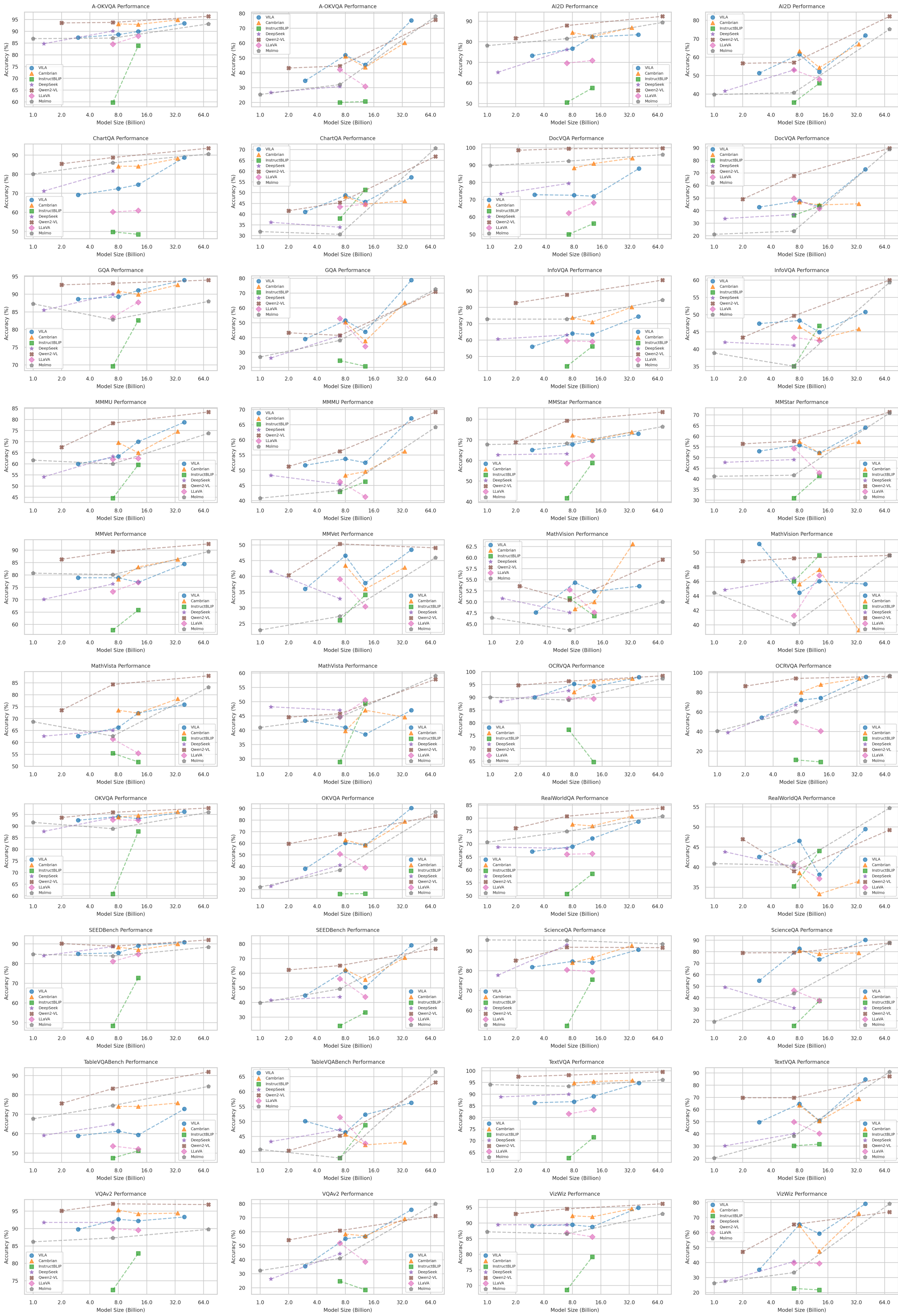


Figure 5: **Model performance and scaling analysis on *NegVQA* across different VLM families and datasets.** For each of the 20 subsets in *NegVQA*, we present scaling curves for both the original non-negated dataset and the negated dataset from left to right, resulting in a total of 40 figures.