SYM-AM-25-425

Excerpt from the Proceedings of the Twenty-Second Annual Acquisition Research Symposium and Innovation Summit

Volume III

Exploring Visual Question Answering Capabilities of Multi-Modal Large Language Models with Model Based Systems Engineering Models

Published: May 5, 2025

The research presented in this report was supported by the Acquisition Research Program at the Naval Postgraduate School.

To request defense acquisition research, to become a research sponsor, or to print additional copies of reports, please contact any of the staff listed on the Acquisition Research Program website (www.acquisitionresearch.net).

ACQUISITION RESEARCH PROGRAM
DEPARTMENT OF DEFENSE MANAGEMENT
NAVAL POSTGRADUATE SCHOOL

Exploring Visual Question Answering Capabilities of Multi-Modal Large Language Models with Model Based Systems Engineering Models

Ryan Longshore—is a 20-year veteran of the defense and electric utility industries. At NIWC LANT, he leads teams developing and integrating new technologies into Navy command centers. He is involved in the Navy's digital engineering transformation with a focus on model-based systems and model-based engineering. Ryan holds a BS in electrical engineering (Clemson) and an MS in systems engineering (SMU), and is pursuing a PhD in systems engineering at the Naval Postgraduate School. He is a South Carolina Registered Professional Engineer and an INCOSE CSEP, and holds the OMG SysML Model Builder Fundamental Certification. [ryan.longshore@nps.edu]

Ryan Bell—is an 9-year experienced engineer in the defense industry. In his current role at Naval Information Warfare Center Atlantic (NIWC LANT), Ryan provides modeling and simulation expertise to a variety of programs for the Navy and USMC. He specializes in simulating communication systems in complex environments and is an advocate for the use of digital engineering early in the systems engineering life cycle. Ryan earned a BS in electrical engineering from Clemson University and a MS in electrical engineering from Clemson University with a focus on electronics, and is currently pursuing his PhD in systems engineering at the Naval Postgraduate School. He is a South Carolina registered Professional Engineer (PE), published author, and teacher. [ryan.bell@nps.edu]

Raymond Madachy, PhD—is a Professor in the Systems Engineering Department at the Naval Postgraduate School. His research interests include systems engineering tool environments for digital engineering, modeling and simulation of systems and software engineering processes, generative AI, and system cost modeling. He has developed cost estimation tools for systems and software engineering, and created the Systems Engineering Library (se-lib). His books include Software Process Dynamics, What Every Engineer Should Know about Modeling and Simulation, What Every Engineer Should Know about Python, and he co-authored Software Cost Estimation with COCOMO II and Software Cost Estimation Metrics Manual for Defense Systems. [rjmadach@nps.edu]

Abstract

The continued advancement of large language models (LLMs) has unlocked new opportunities for systems engineering particularly in the field of visual question answering (VQA). Multi-modal LLMs are capable of processing both textual and graphical inputs, allowing them to interpret the graphical elements of model-based systems engineering (MBSE) models alongside accompanying textual descriptions. This paper explores the capabilities of multi-modal LLMs in understanding and interpreting Systems Modeling Language (SysML) v1 block definition diagrams (BDDs). BDDs are visual diagrams that formally capture a system's structural elements, properties, relationships, and multiplicities.

We evaluate both proprietary and open-source multi-modal LLMs using a curated dataset of SysML BDDs and associated multiple-choice question set designed to assess LLM performance at the first two levels of Bloom's Taxonomy, Remember and Understand. We also analyzed the effect of model size on accuracy. The results provide insights into which current LLMs are able to natively interpret SysML BDD syntax which informs future research aimed at enhancing systems modeling processes with AI agents.

Introduction

The integration of artificial intelligence (AI) into Model-Based Systems Engineering (MBSE) processes presents significant opportunities for improving model comprehension, validation, and support activities. Multi-modal large language models (LLMs) are capable of processing both textual and graphical inputs and have expanded the potential for automating the interpretation of system modeling language (SysML) v1 models. Block Definition Diagrams

(BDDs) are key elements of SysML v1 models, serving as a foundational representation of system structure, properties, and relationships (OMG, 2019).

Despite the rapid evolution of LLMs, their ability to accurately interpret SysML artifacts remains largely unexplored. Existing evaluations of multi-modal LLMs have primarily focused on images or general diagrammatic reasoning, rather than domain-specific graphical languages such as SysML (Antol et al., 2015; Ishmam et al., 2024; Lin et al., 2014). This gap limits the current understanding of LLMs' effectiveness in supporting engineering workflows that rely on formal SysML model interpretation.

This paper addresses this gap by evaluating the performance of contemporary multi-modal LLMs in interpreting SysML v1.x BDDs. We develop a curated dataset of BDDs and design a multiple-choice question set aligned with the first two levels of Bloom's Taxonomy. The evaluation examines both proprietary and open-source LLMs, analyzing their capabilities across models of varying sizes. The findings offer empirical insights into the strengths and limitations of current LLMs in understanding formal systems modeling artifacts and inform future research on enhancing AI-driven support for MBSE practices.

Background and Related Research

Visual Question Answering

Visual question answering (VQA) is a field of AI research focused on answering textual questions using image(s) as contextual input (Antol et al., 2015). Responses can be binary (yes/no), multiple choice, or open-ended. Early VQA methods combined computer vision (CV) feature extraction and natural language processing (NLP) machine learning (ML) techniques to generate answers (Ishmam et al., 2024). The introduction of attention mechanisms such as stacked attention networks and dynamic memory networks enabled multi-step reasoning in VQA tasks (Xiong et al., 2016; Yang et al., 2016). Large Visual Language Models (LVLMs) such as ViLBERT and VisualBERT further advanced the field by incorporating pretraining techniques and transformer architecture to increase model performance (Li et al., 2019; Lu et al., 2019).

The emergence of multi-modal LLMs transformed VQA by enabling unified reasoning over text and images. Models like Flamingo and PaLI demonstrated that scaling vision-language pretraining yields strong few-shot VQA capabilities (Alayrac et al., 2022; Chen et al., 2023). BLIP-2 (Li et al., 2023) further streamlined this approach by efficiently connecting frozen pre-trained image encoders and LLMs (Li et al., 2023). OpenAI's GPT-4 represented a shift toward general-purpose multi-modal reasoning achieving similar performance to text only inputs without VQA-specific architectures (OpenAI et al., 2024). These advancements have moved VQA from specialized models toward foundation models with broad applicability across engineering and scientific tasks.

VQA Benchmarks

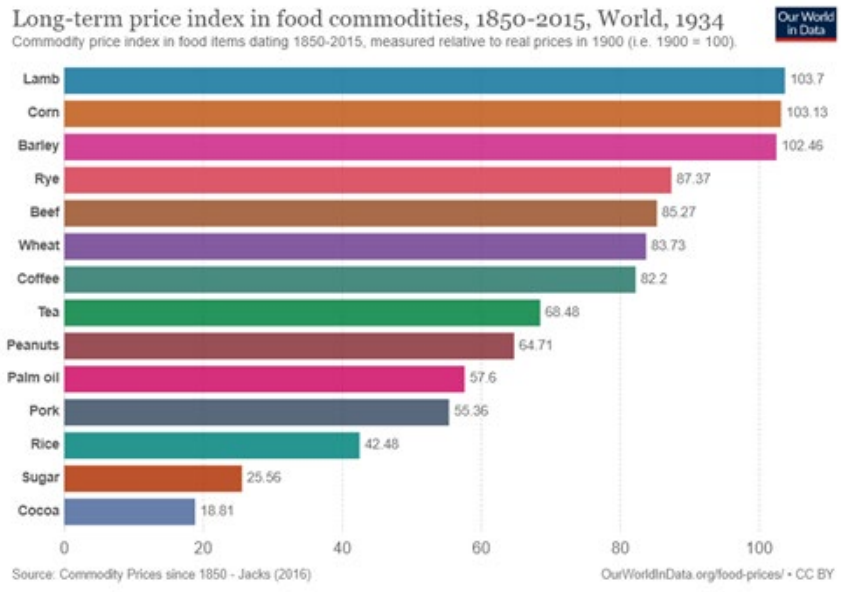
A variety of datasets have been developed to benchmark VQA capabilities. The Dataset for Question Answering on Real-Work images (DAQUAR) was one of the first largely used VQA benchmarks and was a modern attempt at a "visual Turing test" (Malinowski & Fritz, 2015). Microsoft's Common Objects in Context (COCO) dataset introduced a large dataset where each image was provided as a raw image and then a segmented image with highlighted objects (Figure 1) that enabled benchmarking for tasks such as counting (Lin et al., 2014). The VQA-2.0 dataset balanced the VQA-1.0 dataset by collecting complementary images for each question ensuring that each question could be applied to different images and yield different answers (Antol et al., 2015; Goyal et al., 2019).

(a) (b)

Figure 1. Sample Image from COCO Dataset (*COCO - Common Objects in Context*, n.d.)
(a) Original Image, (b) Segmented Image Displaying Overlay for Fire Hydrants and Vehicles

While datasets such as DAQUAR, COCO, and VQA 2.0 addressed general VQA questions, they also highlighted the need for application specific datasets such as chart and diagram specific datasets. These chart and diagram specific datasets sought to address recalling and synthesizing data from the chart using methods such as optical character recognition (OCR), interpreting numerical data contained in a chart, and understanding of different chart structures (Kafle et al., 2020).

Chart and figure specific datasets continue to evolve along with new methods to improve complex reasoning (Srivastava & Sharma, 2024). Data visualization question answering (DVQA) is one of the early chart datasets specifically focused on bar charts (Kafle et al., 2018). When introducing the DVQA dataset, Kafle et al. showed that VQA methods were not effective at recalling or synthesizing data related to bar charts and proposed new methods for chart specific VQA. Also introduced in 2018, FigureQA expanded charts types to include line plots, dot-line plots, and pie charts in addition to bar charts and proposed a Relation Network method (Kahou et al., 2018). More recent datasets such as ChartQA introduce complex reasoning questions that require logical and arithmetic calculations (Masry et al., 2022). A sample from the ChartQA dataset shown in Figure 2 demonstrates the increased complexity of questions. Answering the questions requires the number of bars in the chart, analyzing their labels for relevance (is it a food or not), and then combining those two pieces of information to determine the correct answer.



Question 1 in the ChartQA 'test' dataset:

Q: How many food items are shown in the bar graph?
A: 14

Figure 2. ChartQA Sample Question and Associated Image (*Lmms-Lab/ChartQA · Datasets at Hugging Face*, 2024)

SysML v1.6 BDDs

In SysML v1.6, "the BDD is used to define blocks in terms of their features, and their structural relationships with other blocks" (Friedenthal et al., 2011). While a BDD can convey many types of information about blocks and their relationships, this paper focuses on the following parts of the BDD as described in *SysML Distilled* (Delligatti, 2014):

- Blocks are fundamental modeling elements that represent system components, subsystems, or other concepts (e.g., actors). They can define both structural and behavioral features.
- Properties are attributes owned by a block that define the internal structure and characteristics.
    - Part properties represent a block's internal structure and are used to model composition.
    - Reference properties represent a relationship to an external structure and are used to show dependency on another block.
    - Value properties represent a quantitative or descriptive attribute of a block (e.g., speed in miles per hour, length in inches)
- Relationships convey composition, abstraction, connection, or dependencies between model elements.
    - Composite associations convey structural decomposition and are denoted by filled in diamonds.

    - Reference associations convey a connection or dependency between two blocks. They may also be shown as reference properties.
    - Generalizations convey inheritance between elements and are denoted by unfilled triangles. The generalized element is known as the supertype while the more specialized element is known as the subtype.
- Multiplicity is a constraint specifying the number of allowable instances, such as one-to-one (1) and one-to-many (1..*). Multiplicity can also be used to model optional components (0..1, 0..*).

Representing the LLM Cognitive Process with Bloom's Revised Taxonomy

Bloom's taxonomy is a hierarchical model of cognition widely used in education to classify learning objectives (Bloom et al., 1956). Bloom's revised taxonomy specifies six cognitive process levels: Remember, Understand, Apply, Analyze, Evaluate, and Create (Krathwohl, 2002). In addition to human cognition, recent research has extended Bloom's revised taxonomy to LLMs.

A recent study analyzing the alignment of existing LLM benchmarks to Bloom's revised taxonomy found that most benchmarks adequately assess the "Remember" and "Understand" levels but do not comprehensively address all six cognitive levels (Huber & Niklaus, 2025). Although "Remember" and "Understand" represent the lowest levels of cognition, LLMs do not always perform the highest at these levels. In a mixed-methods study examining ChatGPT's performance on psychosomatic medicine examination questions, researchers observed that GPT-4 exhibited notable deficiencies in these two levels, with 29 errors in "Remember" and 23 errors in "Understand" stemming from difficulties in recalling specific details and grasping conceptual relationships (Herrmann-Werner et al., 2024).

Consistent with other evaluation approaches, this study focuses on the first two levels of Bloom's revised taxonomy: Remember and Understand. "Remember" questions are designed to assess recall of information directly from SysML BDDs without requiring synthesis of multiple elements. "Understand" questions assess higher cognitive engagement through summarization and inference tasks. Summarization questions require synthesis of multiple pieces of information from the diagram while inference questions involve drawing conclusions that are not explicitly stated but are logically supported by the diagram's structure and consistent with SysML v1.6 rules.

Methodology

This section describes the methodology for constructing and evaluating a dataset aimed at assessing LLMs' ability to interpret SysML v1.x BDDs. In the absence of existing datasets focused specifically on SysML, a novel dataset was developed to capture both syntactic and semantic understanding of BDDs through structured multiple-choice questions aligned with Bloom's revised taxonomy (Krathwohl, 2002). A set of both proprietary and open source multi-modal LLMs were evaluated against this dataset. LLM inference was conducted using GPU-accelerated environments and automated through scripting to ensure consistency and reproducibility. The evaluation process culminated in a human as judge assessment of LLM responses where the human judge was a practicing systems engineer.

Dataset Generation

While there are several datasets focused on VQA and diagrams in particular, there are no datasets specifically focused on SysML v1.x. Therefore, the dataset for this analysis was generated by systems engineers with experience in both systems modeling and benchmark

dataset generation. It consists of a curated set of SysML BDDs and associated multiple choice questions. The dataset was exclusively human-generated with no synthetic content.

The dataset consists of 80 questions. Generated questions cover four key concepts from SysML v1.x BDDs: Blocks, Properties, Relationships, and Multiplicity. The difficulty of the generated questions is evenly distributed across the remember and understand levels of Bloom's revised taxonomy. Table 1 details the distribution of questions across both Bloom's Taxonomy and BDD concept.

Table 1. Distribution of Questions

	Blocks	Properties	Relationships	Multiplicity
Remember	10	10	10	10
Understand	10	10	10	10

The dataset follows a syntax common to multiple choice question datasets with some minor modifications to incorporate additional fields such as diagram reference, SysML Concept, and Bloom Taxonomy Category as shown in Table 2. This syntax will allow the dataset to be easily expanded to more diagram types and potentially be incorporated as an extension into systems engineering specific benchmarks such as SysEngBench (Bell, 2024).

Table 2. Dataset Fields

Field	Data Format	Description
QuestionID	Integer	Unique identifier for each question
BDDConcept	Enumeration	One of four options: Blocks, Relationships, Properties, Multiplicity
BloomCategory	Enumeration	One of two options: Remember, Understand
Diagram	String	File name of the associated SysML BDD
Question	String	Text of the multiple choice question
ChoiceA	String	Text for choice A
ChoiceB	String	Text for choice B
ChoiceC	String	Text for choice C
ChoiceD	String	Text for choice D
Answer	String	Correct Answer: ChoiceA, ChoiceB, ChoiceC, ChoiceD

A camera specification BDD from the dataset is shown in Figure 3. This diagram incorporates blocks, value properties, a generalization relationship (denoted by the unfilled triangle), and other elements such as requirements and value types. Two sample questions based on this diagram are shown below. Note that the rationale field is included as a courtesy explanation to the reader as to why the answer is correct, but is not included in the dataset.

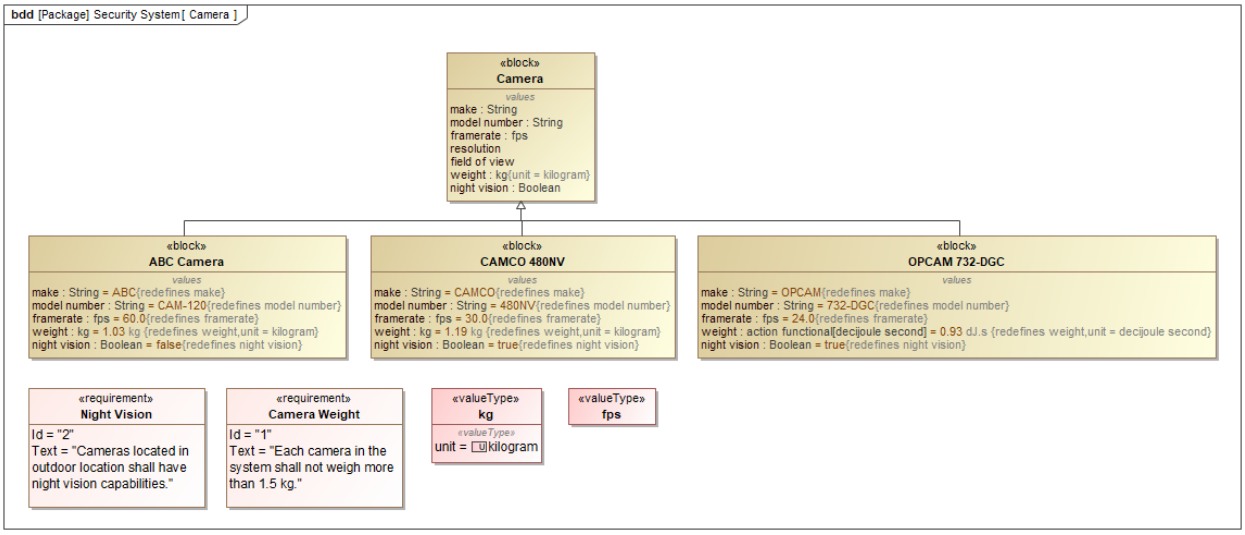


Figure 3. Camera Specification Diagram

A sample “Remember” question related to the BDD properties concept is:

What is the custom value type defined for framerate?

a) string

b) kg

c) boolean

d) fps

Correct Answer: d

Rationale: *The framerate property is typed by the fps property in the ‘framerate : fps’ value property definition.*

A sample *understand* question related to the BDD properties category is:

How many value properties are there for each camera?

a) 15

b) 22

c) 5

d) 7

Correct Answer: d

Rationale: *The generalized camera block contains seven value properties that are inherited by each camera. Each specific camera block shows the five value properties that are re-defined, but not the inherited properties that are not re-defined.*

LLM Selection

A variety of open source and proprietary models easily accessible to practicing engineers were selected for this paper. The open source models were selected as they are the multi-modal vision models available from the widely used Ollama library as of April 2025

(*Ollama*, n.d.). ChatGPT-4o and Sonnet-3.7 were selected as they are widely available proprietary models. The dataset was evaluated against the following models:

- baklava: 7B
- gemma3: 4B, 12B, and 27B Variants
- granite3.2-vision: 2B
- llama3.2-vision: 11B and 90B Variants
- llava: 7B, 13B and 34B Variants
- llava-llama3: 8B
- minicpm-v: 8B
- mistral-small3.1: 24B
- moondream: 1.8B
- OpenAI chatgpt-4o
- Anthropic sonnet-3.7

Dataset Evaluation

To evaluate the dataset using Ollama models, a virtual GPU pod instance was provisioned on RunPod utilizing an NVIDIA A40 GPU. Ollama was installed on this virtual pod following the guidelines provided in the RunPod documentation (*Set up Ollama on Your GPU Pod | RunPod Documentation*, 2025). The selected LLMs were then loaded into the GPU pod via the ollama *run* and *pull* commands. A Jupyter Notebook was deployed within the same pod to facilitate the evaluation process. The question set formatted as a CSV file along with the corresponding images was uploaded to the notebook environment. A Python script was developed to automate the process of asking questions to the LLMs and capturing their responses. The script iterated through each question in the question set, submitted each prompt to the LLM under evaluation, and recorded the generated answers. The outputs were then written to a CSV file for analysis. This workflow is detailed in Figure 4.

The same dataset was used to evaluate the chatgpt-4o and sonnet-3.7 models. However, instead of using custom scripts and dedicated GPUs, the ChatGPT (*ChatGPT*, n.d.) and Claude (*Claude*, n.d.) websites were utilized to ask the LLMs questions.

LLMs do not explicitly know they should answer a multiple choice question with a one character response. Therefore the question was asked in the following format:

You are an automated system that answer multiple choice questions and only outputs one of four letters: A, B, C, or D. Given the following question and four answer choices, respond with ONLY the letter of the best answer. This will be A, B, C, or D. Do not explain your answer. Do not say anything else. Use the image as context for your answer.

Question: {question}

A. {option_a}

B. {option_b}

C. {option_c}

D. {option_d}

There are several methods to compare the model answers to the correct answers including LLM as a judge and human as a judge. LLM as a judge refers to the use of LLMs as automated judges for evaluating other LLMs on open-ended tasks where traditional benchmarks may be insufficient (Zheng et al., 2023). However, due to limitations in dataset size (80 questions), model coverage (18 models), and the fact that LLM judging focuses on evaluating the final answer rather than the reasoning process behind it, human as a judge is employed for the final assessment.

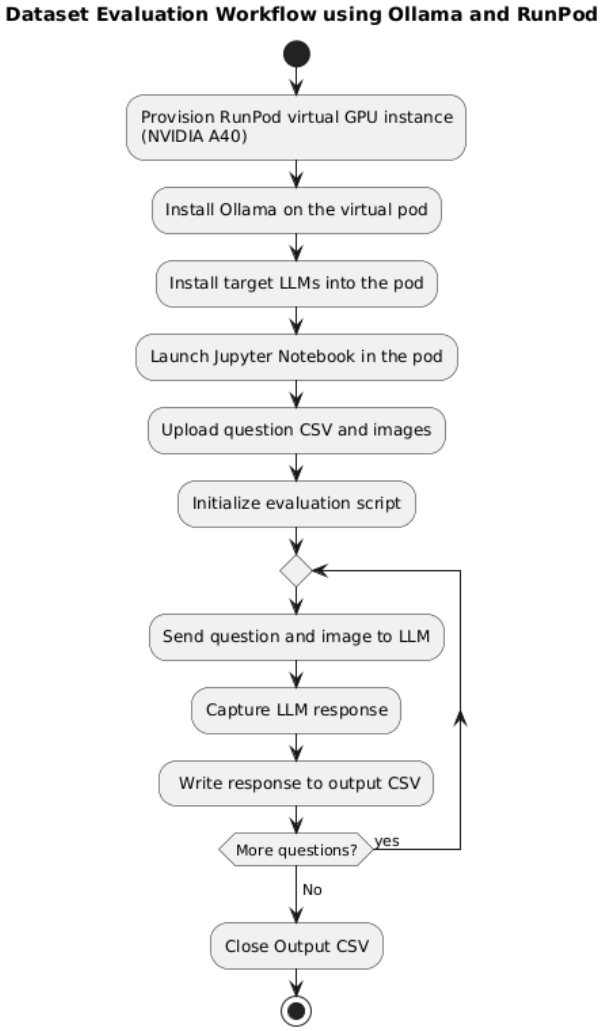


Figure 4. Dataset Evaluation Workflow

Results, Discussion, and Limitations

Each LLM's responses were scored against the correct multiple-choice answers to evaluate accuracy. Accuracy is defined as the percentage of questions answered correctly. The dataset was designed to assess both syntactic and semantic understanding of SysML BDDs covering a balanced distribution across four modeling concepts and two levels of Bloom's revised taxonomy. The results presented below compare overall model performance, analyze trends relative to model size, and break down accuracy by cognitive level and SysML concept.

The overall performance of each LLM is shown in Figure 5. Proprietary LLMs are denoted by orange bars while open source LLMs are denoted by blue bars. Although proprietary models secured two of the top three scores, the open-source model mistral-small3.1, a 24B model, outperformed Sonnet-3.7 while falling short of GPT-4o. Given that each multiple-choice question included four possible answers, the expected accuracy from random guessing across all 80 questions is 25%. Bakllava, a 7B model, demonstrated the lowest performance and was the only model that failed to exceed the random guessing baseline.

The scatter plot in Figure 6 compares LLM accuracy to model size. It is important to note that the size of GPT-4o and Sonnet-3.7 is not publicly available information. There are several estimates of around 200 billion parameters, but those estimates have not been confirmed by either OpenAI or Anthropic. A correlation coefficient of 0.65 indicates a moderate relationship between LLM size and accuracy. However, mistral-small3.1 (24B) outperforms three larger open source models as well as Sonnet-3.7. Despite being the second smallest model, granite3.2-vision (2B) outperforms 10 larger models. These observations suggest that factors beyond parameter count, such as training data and/or methods, influence performance.

The grouped bar chart in Figure 7 visualizes accuracy by Bloom's revised taxonomy category. Most LLMs perform better on "Remember" tasks than on "Understand" questions with GPT-4o correctly answering all "Remember" questions. Two LLMs performed slightly better on "Understand" questions. These results indicate the LLMs' ability to recall information from a diagram is greater than the ability to synthesize multiple pieces of information or bring in additional context not explicitly stated in the BDD.

The multi-series bar chart shown in Figure 8 breaks down performance across the four core SysML v1.6 BDD concepts: Blocks, Relationships, Properties, and Multiplicity. The results reveal notable variation across concepts, with most models performing best on Blocks and worst on Relationships or Multiplicity, highlighting uneven conceptual understanding amongst LLMs.

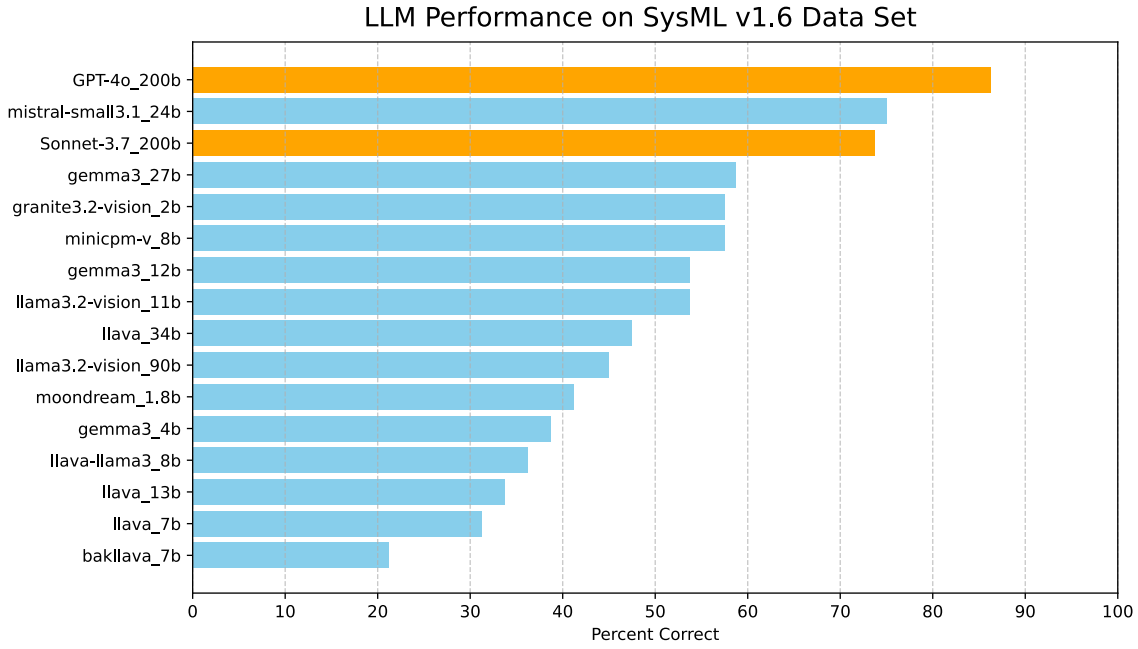


Figure 5. Overall Performance

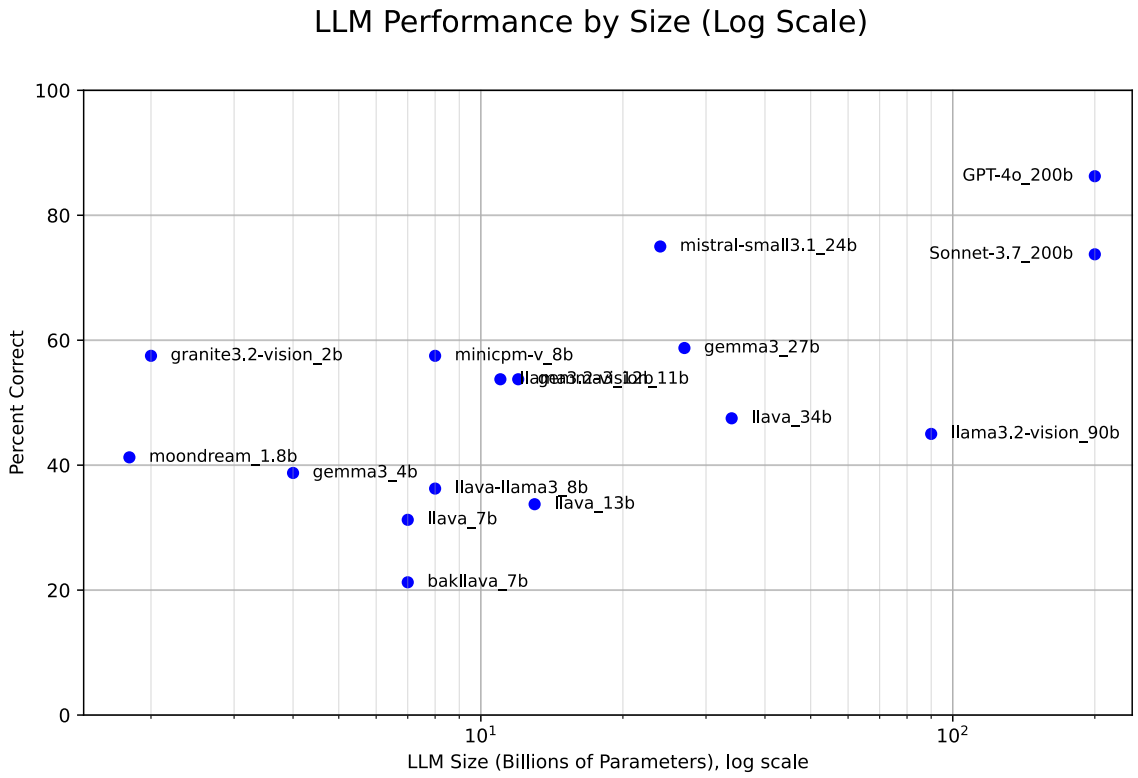


Figure 6. Performance by LLM Size

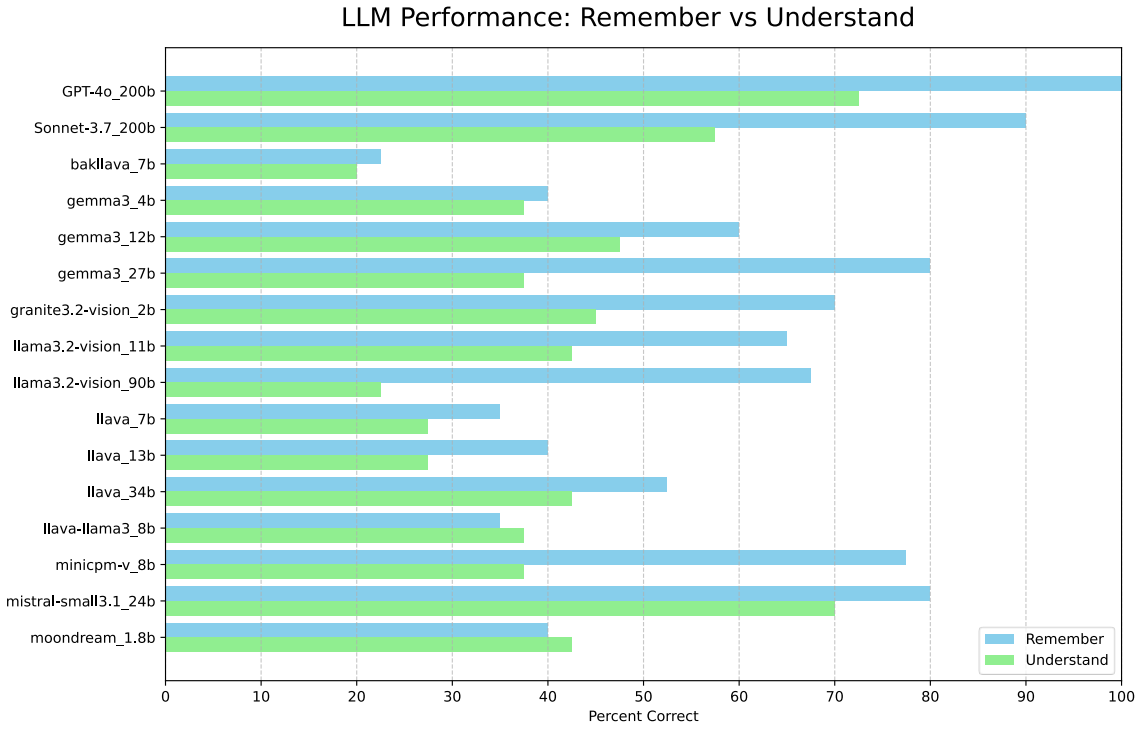


Figure 7. Performance by Bloom's Taxonomy Category

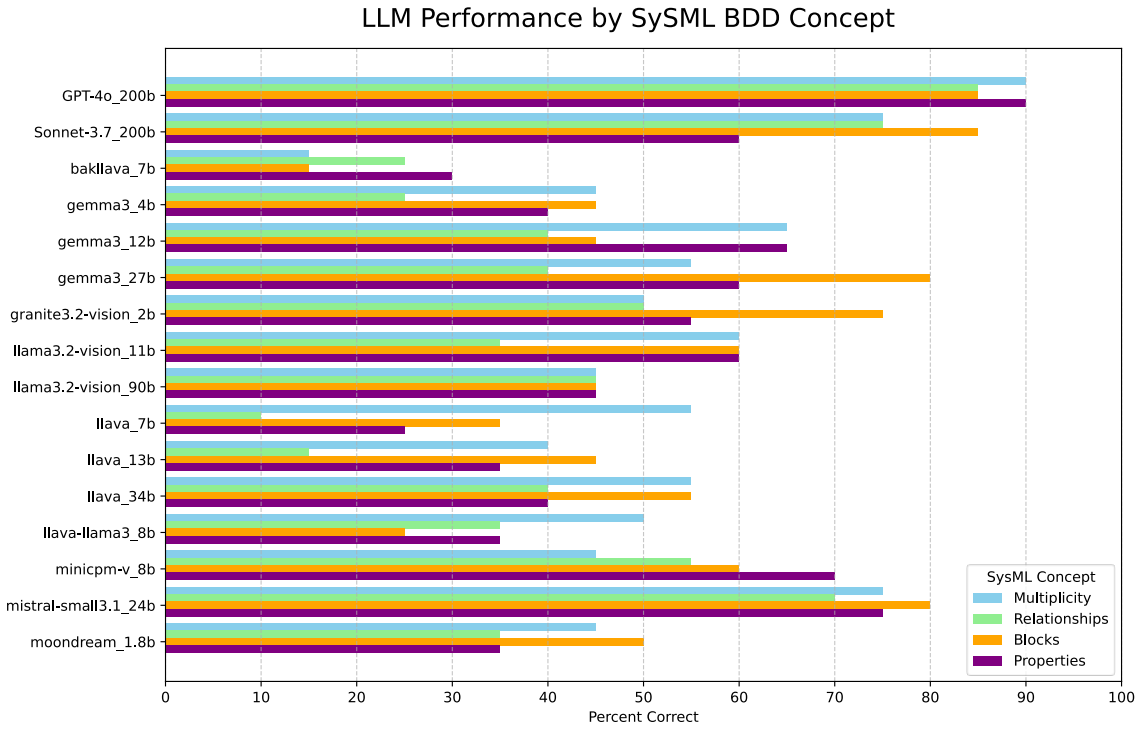


Figure 8. Performance by BDD Concept

Future Work

This study is an initial exploration of LLM performance on SysML v1.x BDDs using a curated dataset of multiple-choice questions. While the current dataset is balanced across four core BDD concepts and two levels of Bloom's revised taxonomy ("Remember" and "Understand"), future research can extend the depth and breadth of this analysis.

Future studies could focus on expanding the dataset in three different ways:

- Expansion of the dataset to include more questions and images. This could increase the robustness of the evaluation and potentially increase the statistical significance of the results.
- Incorporation of additional SysML v1.x diagram types beyond BDDs such as Internal Block Diagrams (IBDs), Activity Diagrams, and Sequence Diagrams would provide a more comprehensive benchmark to evaluate the extent to which LLMs can generalize across different visual and semantic structures in systems modeling. This would also increase the number of multiple-choice questions per Bloom's revised taxonomy level to improve statistical robustness and reduce sensitivity to specific wording or diagram features.
- Expansion of the dataset to include higher levels of Bloom's revised taxonomy, such as "Apply," "Analyze," "Evaluate," and potentially even "Create" could give a more holistic view of LLM capabilities. By incorporating these more complex cognitive tasks, future studies can investigate whether LLM performance declines as tasks become more abstract and cognitively demanding.

This study identified several LLMs that may be promising candidates for techniques such as Retrieval-Augmented Generation (RAG) to improve accuracy. Applying RAG could allow models to draw from relevant SysML documentation or design patterns to enhance their

question answering abilities. Future experiments could explore the impact of RAG on accuracy particularly in handling the more difficult “Understand” questions or tasks at higher levels or Bloom’s revised taxonomy.

Conclusion

This study presents a targeted evaluation of multi-modal LLMs on SysML v1.6 BDDs through a VQA framework. By grounding the analysis in Bloom’s revised taxonomy and assessing both proprietary and open-source models, we provide empirical insights into how LLMs interpret formal, domain-specific systems modeling diagrams. The findings show that while model size moderately correlates with accuracy, other factors also impact LLM performance. Most models demonstrate stronger capabilities in recalling elements (“Remember”) than in synthesizing or inferring information (“Understand”) revealing limitations in semantic comprehension of structured graphical artifacts.

The curated dataset and evaluation framework introduced here lay the groundwork for future research into more advanced cognitive tasks and broader SysML diagram types. As the field progresses, improving model performance through techniques like RAG on domain-specific content holds significant promise. Ultimately, understanding and enhancing how LLMs process systems modeling artifacts is a critical step toward their meaningful integration into MBSE workflows.

Acknowledgements

This work has benefited from the use of generative AI tools including ChatGPT and SciSpace for writing assistance and code development. These tools were employed to enhance conceptual clarity, improve code efficiency, and support literature synthesis. All AI-generated outputs were critically reviewed, refined, and validated to ensure accuracy and alignment with academic integrity. Their contributions were limited to supporting the research process, and final responsibility for the content, analysis, and conclusions remains with the authors.

References

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., … Simonyan, K. (2022). *Flamingo: A visual language model for few-shot learning* (No. arXiv:2204.14198). ArXiv. https://doi.org/10.48550/arXiv.2204.14198

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., & Parikh, D. (2015). VQA: Visual question answering. *2015 IEEE International Conference on Computer Vision (ICCV)*, 2425–2433. https://doi.org/10.1109/ICCV.2015.279

Bell, R. (2024, August). Rabell/SysEngBench \cdot datasets at Hugging Face. In *SysEngBench*. https://huggingface.co/datasets/rabell/SysEngBench

Bloom, B. S., Engelhart, M. D., Furst, E., Hill, W. H., & Krathwohl, D. R. (1956). *Handbook I: Cognitive domain*. David McKay, 483–498.

ChatGPT. (n.d.). Retrieved April 24, 2025, from https://chatgpt.com

Chen, X., Wang, X., Changpinyo, S., Piergiovanni, A. J., Padlewski, P., Salz, D., Goodman, S., Grycner, A., Mustafa, B., Beyer, L., Kolesnikov, A., Puigcerver, J., Ding, N., Rong, K., Akbari, H., Mishra, G., Xue, L., Thapliyal, A., Bradbury, J., … Soricut, R. (2023). *PaLI: A jointly-scaled multilingual language-image model* (No. arXiv:2209.06794). ArXiv. https://doi.org/10.48550/arXiv.2209.06794

Claude. (n.d.). Retrieved April 24, 2025, from https://claude.ai/new

COCO - Common Objects in Context. (n.d.). Retrieved April 20, 2025, from https://cocodataset.org/#explore

Delligatti, L. (2014). *SysML distilled*. Pearson Education.

Friedenthal, S., Moore, A., & Steiner, R. (2011). *A practical guide to SysML: The systems modeling language*. Elsevier Science & Technology. http://ebookcentral.proquest.com/lib/ebook-nps/detail.action?docID=787244

Goyal, Y., Khot, T., Agrawal, A., Summers-Stay, D., Batra, D., & Parikh, D. (2019). Making the V in VQA matter: Elevating the role of image understanding in visual question answering. *International Journal of Computer Vision*, *127*(4), 398–414. https://doi.org/10.1007/s11263-018-1116-0

Herrmann-Werner, A., Festl-Wietek, T., Holderried, F., Herschbach, L., Griewatz, J., Masters, K., Zipfel, S., & Mahling, M. (2024). Assessing ChatGPT's mastery of Bloom's taxonomy using psychosomatic medicine exam questions: Mixed-methods study. *Journal of Medical Internet Research*, *26*, e52113. https://doi.org/10.2196/52113

Huber, T., & Niklaus, C. (2025). LLMs meet Bloom's taxonomy: A cognitive view on large language model evaluations. In O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, & S. Schockaert (Eds.), *Proceedings of the 31st International Conference on Computational Linguistics* (pp. 5211–5246). Association for Computational Linguistics. https://aclanthology.org/2025.coling-main.350/

Ishmam, Md. F., Shovon, Md. S. H., Mridha, M. F., & Dey, N. (2024). From image to language: A critical analysis of visual question answering (VQA) approaches, challenges, and opportunities. *Information Fusion*, *106*, 102270. https://doi.org/10.1016/j.inffus.2024.102270

Kafle, K., Price, B., Cohen, S., & Kanan, C. (2018). DVQA: Understanding data visualizations via question answering. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5648–5656. https://doi.org/10.1109/CVPR.2018.00592

Kafle, K., Shrestha, R., Price, B., Cohen, S., & Kanan, C. (2020). Answering questions about data visualizations using efficient bimodal fusion. *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 1487–1496. https://doi.org/10.1109/WACV45572.2020.9093494

Kahou, S. E., Michalski, V., Atkinson, A., Kadar, A., Trischler, A., & Bengio, Y. (2018). *FigureQA: An annotated figure dataset for visual reasoning* (No. arXiv:1710.07300). ArXiv. https://doi.org/10.48550/arXiv.1710.07300

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory into Practice*, *41*(4), 212.

Li, J., Li, D., Savarese, S., & Hoi, S. (2023). *BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models* (No. arXiv:2301.12597). ArXiv. https://doi.org/10.48550/arXiv.2301.12597

Li, L. H., Yatskar, M., Yin, D., Hsieh, C.-J., & Chang, K.-W. (2019). *VisualBERT: A simple and performant baseline for vision and language* (No. arXiv:1908.03557). arXiv. https://doi.org/10.48550/arXiv.1908.03557

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In D. Fleet, T. Pajdla, B. Schiele, &

T. Tuytelaars (Eds.), *Computer vision—ECCV 2014* (pp. 740–755). Springer International Publishing. https://doi.org/10.1007/978-3-319-10602-1_48

Lmms-lab/ChartQA · Datasets at Hugging Face. (2024, October 4). https://huggingface.co/datasets/lmms-lab/ChartQA

Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). *ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language rasks* (No. arXiv:1908.02265). ArXiv. https://doi.org/10.48550/arXiv.1908.02265

Malinowski, M., & Fritz, M. (2015). *A multi-world approach to question answering about real-world scenes based on uncertain input* (No. arXiv:1410.0210). ArXiv. https://doi.org/10.48550/arXiv.1410.0210

Masry, A., Do, X. L., Tan, J. Q., Joty, S., & Hoque, E. (2022). ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), *Findings of the association for computational linguistics: ACL 2022* (pp. 2263–2279). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.findings-acl.177

Ollama. (n.d.). *Vision models · Ollama Search*. Retrieved April 18, 2025, from https://ollama.com/search

OMG. (2019, November). *OMG Systems Modeling Language version 1.6*. https://www.omg.org/spec/SysML/1.6/PDF

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., … Zoph, B. (2024). *GPT-4 technical report* (No. arXiv:2303.08774). ArXiv. https://doi.org/10.48550/arXiv.2303.08774

Set up Ollama on your GPU Pod | RunPod Documentation. (2025). https://docs.runpod.io/tutorials/pods/run-ollama

Srivastava, S., & Sharma, H. (2024). Deep multimodal relational reasoning and guided attention for chart question answering. *Journal of Electronic Imaging*, *33*(6), 063052. https://doi.org/10.1117/1.JEI.33.6.063052

Xiong, C., Merity, S., & Socher, R. (2016). *Dynamic memory networks for visual and textual question answering* (No. arXiv:1603.01417). ArXiv. https://doi.org/10.48550/arXiv.1603.01417

Yang, Z., He, X., Gao, J., Deng, L., & Smola, A. (2016). *Stacked attention networks for image question answering* (No. arXiv:1511.02274). ArXiv. https://doi.org/10.48550/arXiv.1511.02274

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). *Judging LLM-as-a-judge with MT-Bench and chatbot arena* (No. arXiv:2306.05685). ArXiv. https://doi.org/10.48550/arXiv.2306.05685

Acquisition Research Program
Department of Defense Management
Naval Postgraduate School
555 Dyer Road, Ingersoll Hall
Monterey, CA 93943

www.acquisitionresearch.net