

SoM-1K: A THOUSAND-PROBLEM BENCHMARK DATASET FOR STRENGTH OF MATERIALS

**Qixin Wan^{1,*}, Zilong Wang^{1,*}, Jingwen Zhou¹, Wanting Wang¹, Ziheng Geng²,
Jiachen Liu³, Ran Cao^{1,†}, Minghui Cheng^{2,4,†}, Lu Cheng^{5,†}**

¹College of Civil Engineering, Hunan University, Changsha, 410082, China

²Department of Civil & Architectural Engineering, University of Miami, Coral Gables, FL 33146, USA

³Department of Electrical and Computer Engineering, University of Miami, Coral Gables, FL 33146, USA

⁴School of Architecture, University of Miami, Coral Gables, FL 33146, USA

⁵Department of Computer Science, University of Illinois Chicago, Chicago, IL 60607, USA

*Equal contribution.

†Corresponding authors: rcao@hnu.edu.cn, minghui.cheng@miami.edu, lucheng@uic.edu

ABSTRACT

Foundation models have shown remarkable capabilities in various domains, but their performance on complex, multimodal engineering problems remains largely unexplored. We introduce SoM-1K, the first large-scale multimodal benchmark dataset dedicated to evaluating foundation models on problems in the strength of materials (SoM). The dataset, which contains 1,065 annotated SoM problems, mirrors real-world engineering tasks by including both textual problem statements and schematic diagrams. Due to the limited capabilities of current foundation models in understanding complicated visual information, we propose a novel prompting strategy called Descriptions of Images (DoI), which provides rigorous expert-generated text descriptions of the visual diagrams as the context. We evaluate eight representative foundation models, including both large language models (LLMs) and vision language models (VLMs). Our results show that current foundation models struggle significantly with these engineering problems, with the best-performing model achieving only 56.6% accuracy. Interestingly, we found that LLMs, when provided with DoI, often outperform VLMs provided with visual diagrams. A detailed error analysis reveals that DoI plays a crucial role in mitigating visual misinterpretation errors, suggesting that accurate text-based descriptions can be more effective than direct image input for current foundation models. This work establishes a rigorous benchmark for engineering AI and highlights a critical need for developing more robust multimodal reasoning capabilities in foundation models, particularly in scientific and engineering contexts.

1 INTRODUCTION

Strength of Materials (SoM) or Mechanics of Materials is a cornerstone of engineering, studying how solid objects deform and fail under loads. Solving SoM problems requires seamlessly integrating multimodal information, both textual and visual. For instance, an engineer must analyze the text describing material properties, while simultaneously interpreting diagrams that illustrate the geometry of the structure and its boundary conditions. This ability to reason across modalities is a fundamental engineering skill, yet it remains a significant challenge for AI (Wang et al., 2025).

Project homepage: <https://som-1k.github.io/>

SoM is also a high-stakes domain, where reasoning errors can lead directly to unsafe designs and structural failures. Reliability is therefore not optional but essential: any AI system deployed in this context must meet the same rigorous standards of safety and precision expected of human engineers. These demands make SoM particularly challenging for AI, as models must not only perform accurate calculations but also correctly interpret schematics and integrate them with textual problem statements.

While foundation models have shown strong performance in text-based mathematical reasoning (Cobbe et al., 2021; Ahn et al., 2024; Seßler et al., 2024), they often struggle with specialized vision-language tasks in engineering. The main reason is that existing vision-language models (VLMs), typically trained on natural images, lack the domain-specific knowledge needed to interpret engineering schematics (Doris et al., 2024). To be useful in engineering, AI must evolve to reason from visual information with the same precision as human experts (Hao et al., 2025).

A key obstacle of developing reliable foundation models for SoM is the lack of suitable datasets. Current datasets are poorly aligned with the unique demands of engineering problem-solving (Picard et al., 2023). Text-only datasets omit critical visual cues, while popular multimodal datasets focus on everyday imagery (Schuhmann et al., 2022) and exclude the specialized symbols and physical principles central to engineering. As a result, no standardized benchmark exists yet to evaluate foundation models on authentic, visually rich SoM problems. Existing evaluations largely emphasize conceptual or text-based questions (Marino et al., 2019), neglecting the multimodal reasoning required to arrive at physically grounded solutions (Bakhtin et al., 2019; Yi et al., 2020). Developing such a benchmark is therefore essential to systematically assess models’ capabilities and advance their reliable use in engineering practice.

To bridge this gap, we introduce **SoM-1K**, the first domain-specific multimodal benchmark that integrates text, equations, and engineering diagrams to reflect the authentic reasoning demands of engineering problems. Unlike prior datasets relying primarily on texts (Hendrycks et al., 2021; Wang et al., 2019), SoM-1K captures the multimodal nature of real engineering practice and establishes a standardized platform for rigorous evaluation. Our contributions are threefold: (1) we present the first large-scale multimodal benchmark tailored to mechanics problems, which can be found in the supplementary materials; (2) we systematically evaluate leading foundation models, revealing their current limitations in visual-textual reasoning; and (3) we propose and validate the use of text-based diagram descriptions (DoI) as an effective prompting strategy to reduce reasoning errors. Together, these contributions not only fill a critical gap in AI evaluation resources but also provide actionable insights for developing the next generation of AI systems capable of reliable engineering reasoning.

2 RELATED WORK

Foundation Models in STEM (Science, Technology, Engineering, and Mathematics). The use of Large Language Models (LLMs) in STEM has grown rapidly. Initially, models like Minerva (Lewkowycz et al., 2022) and PaLM (Chowdhery et al., 2022) excelled at solving complex math and physics problems by using techniques like chain-of-thought (CoT) prompting (Wei et al., 2023). This success has expanded into various engineering disciplines, where LLMs assist with design, simulations (Liu et al., 2024), and inverse problems, often by integrating with external tools (Niketani et al., 2025). For instance, LLMs are being applied in bridge engineering to interpret and process the vast amount of unstructured data found in inspection reports, transforming it into structured, actionable insights for decision support (Kumar & Agrawal, 2025). The development of VLMs has also been crucial, allowing models to interpret diagrams and schematics (Picard et al., 2024), a core part of engineering education. These VLMs are now used in educational settings to provide interactive, step-by-step guidance by analyzing visual inputs (Bewersdorff et al., 2025; Scarlatos et al., 2025).

Benchmarking in Engineering Domains. Existing benchmarks in engineering domains have highlighted the challenges faced by AI models in interpreting and reasoning over technical diagrams and textual information. For instance, the DesignQA benchmark evaluates VLMs on tasks involving engineering documentation, CAD images, and textual design requirements, revealing significant gaps in model performance when both visual and textual information are required (Doris et al., 2024). Similarly, the EEE-Bench benchmark assesses VLMs on practical engineering tasks in electrical and electronics engineering, demonstrating that current models often struggle with complex

visual and textual integration, achieving average performance ranging from 19.48% to 46.78% (Li et al., 2025b). These studies underscore the necessity for benchmarks that rigorously evaluate AI models’ abilities to handle multimodal engineering problems, including the integration of schematic diagrams and textual descriptions.

AI Assistance in Mechanics of Materials. In the field of mechanics of materials, several projects have explored the use of AI and LLMs (Tian & Zhang, 2024; Buehler, 2023; Ni & Buehler, 2023; Liu et al., 2025). For instance, the AutoGen (Tian & Zhang, 2024) aimed to presents a framework where multiple LLM-based agents collaborate to solve mechanics problems using the Finite Element Method. The MechAgent (Ni & Buehler, 2023) introduced a novel multi-agent paradigm where a team of AI agents with specialized roles collaboratively automates the process of solving complex mechanics tasks.

To the best of our knowledge, however, no multimodal benchmark study has yet evaluated the reasoning capabilities of foundation models in solving mechanics problems.

3 THE SOM-1K DATASET

3.1 BACKGROUND IN SOM








SoM is a fundamental branch of engineering that studies how solid objects respond to external forces, such as tension, compression, torsion, and bending. Problems in this domain typically focus on analyzing why materials fail, a fundamental concern underlying nearly all engineered systems, from bridges and aircraft to robots and microchips. For this reason, SoM is a core subject in civil, mechanical, aerospace, and materials engineering curricula worldwide, and accurate problem-solving in this domain underpins real-world engineering design and decision-making. Hence, SoM provides an ideal domain for evaluating foundation models’ reasoning capabilities, as it requires the integration of physical principles, mathematical formulations, and the logical application of boundary conditions, paralleling the forms of reasoning demanded in complex coding and scientific problem-solving.

3.2 SCOPE OF THE DATASET

Our multimodal benchmark dataset, **SoM-1K**, is designed to evaluate AI models on authentic mechanics problems. It includes the three fundamental problem types: axial loading (bars), torsion (shafts), and bending (beams and frames) (Hibbeler, 2012). SoM-1K spans a wide range of calculation tasks, including computation of internal forces, stresses, strains, and deformations, diagram construction, and design-oriented optimizations.

Problems were carefully selected from widely-used university textbooks (Sun et al., 2009; Huang, 2009; Dai, 2015; Ma, 2011; Hibbeler, 2012; Gere & Goodno, 2009; Guo & Liu, 2010) and advanced mechanics competitions, ensuring a hierarchical dataset encompasses both routine exercises and more challenging tasks. We consolidate all source materials into PDF format, with textbooks scanned from physical copies and competition problems obtained from official exam websites (Chinese Society of Theoretical and Applied Mechanics & Zhou Peiyuan Foundation, 2025).

Table 1: Statistics of dataset composition in SoM-1K.

Category	Quantity	Proportion
Classified by deformation modes		
 Axial loading (bars)	201	18.87%
 Torsion (shafts)	137	12.86%
 Bending-I (beams)	630	59.15%
 Bending-II (frames)	54	5.07%
 Integrated tasks	43	4.04%
Overall	1065	100%
Classified by statical indeterminacy		
 Statically determinate (easy)	917	86.10%
 Statically indeterminate (hard)	148	13.90%

In total, SoM-1K comprises 1,065 annotated problems, summarized in Table 1, categorized into five groups based on structural components and loading conditions: (1) Axial loading (bars), (2) Torsion (shafts), (3) Bending-I (beams), (4) Bending-II (frames), and (5) Integrated tasks. Integrated problems, sourced from mechanics competitions, require multi-concept reasoning, combining static

analysis with dynamic concepts such as vibration, impact, and rigid-body motion. Example problems from each category are provided in Figure 7 (Appendix A).

3.3 COMPONENTS OF THE DATASET

An illustrative example of the dataset structure is shown in Figure 1. Each problem consists of four standardized components:

- (1) **Problem Statement (PS):** A concise textual description of the problem that specifies the given information and the quantity or outcome to be determined.
- (2) **Schematic Diagram (Image, I):** A graphical representation of the structure or an object, provided in image format. Throughout this work, the term *Image* refers to such schematic diagrams.
- (3) **Description of the Image (DoI):** Expert-validated text describing schematics (e.g., geometry, boundary conditions), providing a precise representation of visual information for evaluating model performance.
- (4) **Ground Truth (GT):** The correct solution to the problem, including equations, reasoning steps, and final answers.

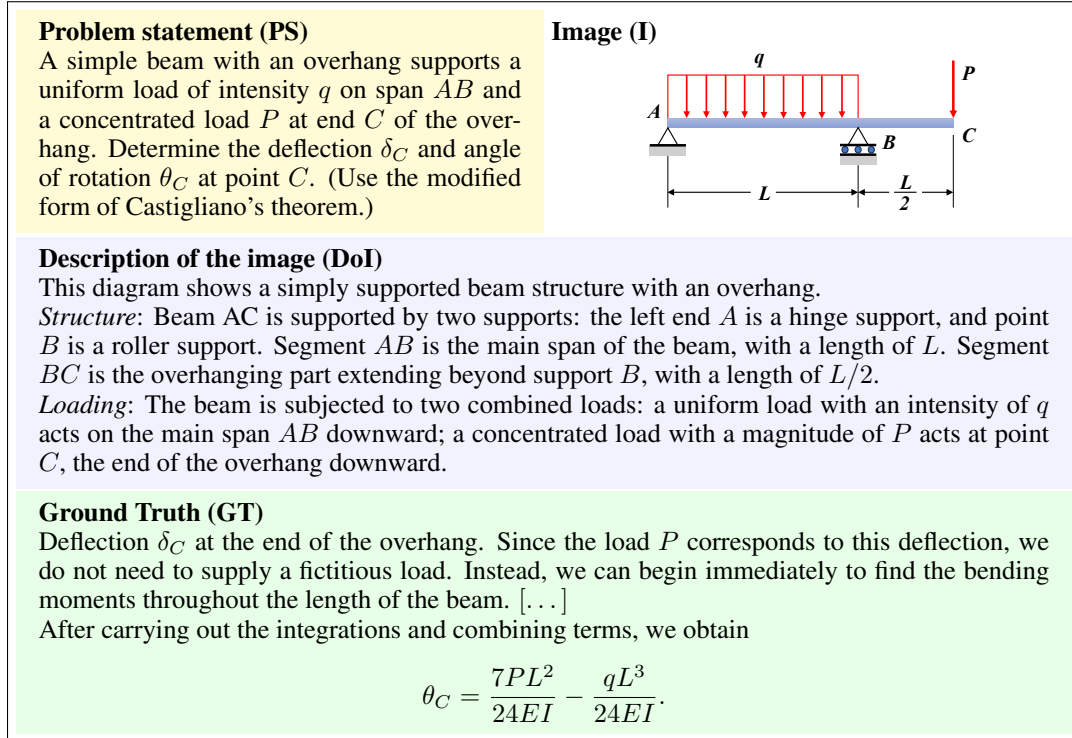


Figure 1: An illustrative example of the dataset structure : problem statement (PS), image (I), description of the image (DoI), and ground truth (GT).

Our workflow began with scanned files of textbooks and problem sets, followed by a two-step pre-processing pipeline. First, schematic diagrams were manually extracted and stored as PNG images. Second, textual content, including PS and GT, was extracted using Doubao (ByteDance, 2025) Optical Character Recognition (OCR). If the GT includes internal force diagrams (Hibbeler, 2012) or other elements that cannot be extracted via OCR, the annotation team manually supplement the description of these diagrams. The extracted text was then carefully reviewed and manually refined to correct OCR errors, ensuring accurate and high-quality representations. The annotation team includes experienced researchers and educators in structural engineering and mechanics of materials, including a PhD candidate, two lecturers, and four teaching assistants.

During preliminary testing, we observed that foundation models struggled to process LaTeX-formatted expressions in batch inference. To mitigate this, we employed the DeepSeek-V3-0324

API (SiliconCloud, 2025) to convert all LaTeX equations into natural-language descriptions, thereby providing consistent textual representations for model inputs.

3.4 DOI ANNOTATION PROCESS

The DoI is derived from the PNG schematics (see Figure 2). Each image is first processed by the Doubao (ByteDance, 2025) VLM, which generates an initial textual description capturing key aspects of the structural diagrams, including geometry, loading conditions, and boundary conditions. These auto-generated descriptions are then carefully reviewed and refined by the annotation team to correct errors and incorporate missing information critical for problem-solving.

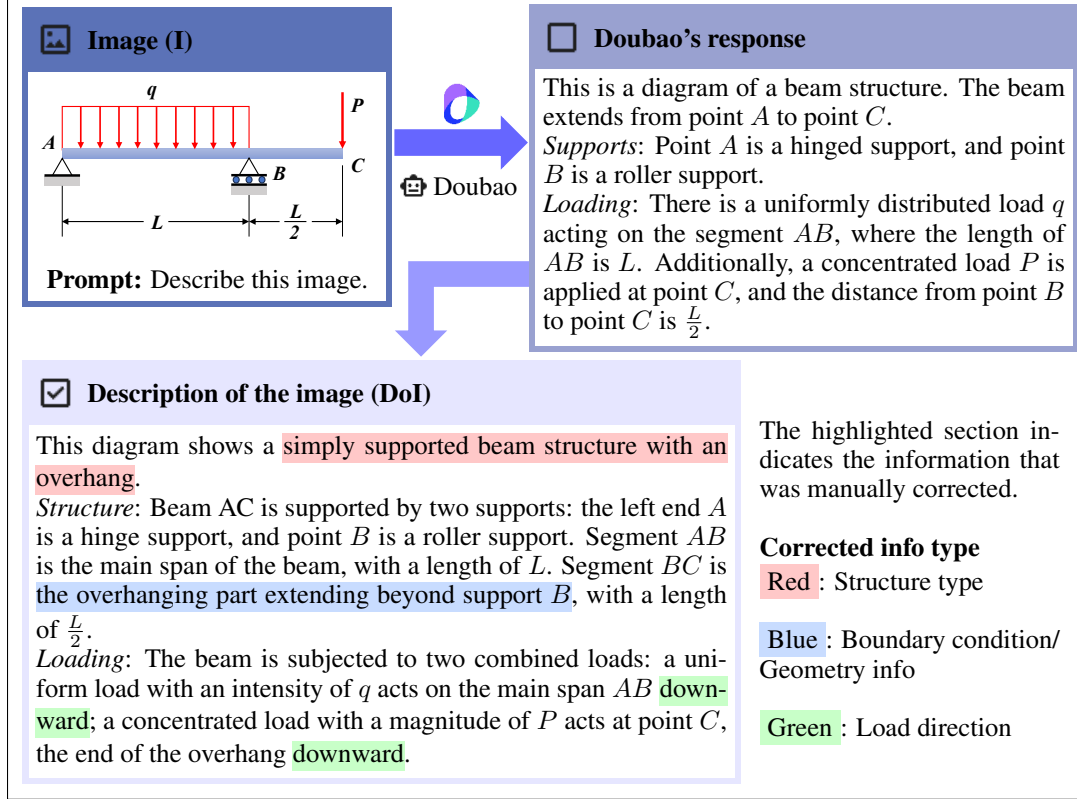


Figure 2: Workflow illustrating the process from the input image to the DoI: an image is first processed by the Doubao VLM, which generates an initial description of the image. This response is then carefully reviewed and corrected by human experts to produce the final DoI. (For improved readability of this colored figure, please refer to the digital version of the paper.)

It is important to note that our DoI is fundamentally different from a typical CoT prompt (Wei et al., 2023). The DoI is designed to only describe the information visually present in the image and does not provide any additional insights or step-by-step reasoning to help the model solve the problem. This clear distinction allows us to isolate and measure the specific impact of descriptive image information on the model’s performance.

4 EVALUATION

4.1 MODELS SELECTED

We evaluate eight representative foundation models on our collected dataset. To ensure a diverse representation of the current landscape, we include both closed-source models (e.g., GPT-4o (OpenAI, 2024), Qwen-plus (Alibaba Cloud, 2025a), Qwen-VL (Alibaba Cloud, 2025b), GPT-3.5 (OpenAI,

2023) and Doubao (ByteDance, 2025)) and leading open-source models (e.g., Llama-70B (Meta AI, 2024), GPT-oss-120b (OpenAI, 2025) and DeepSeek-R1 (DeepSeek, 2025)).

Among them, LLMs include GPT-oss-120b, Qwen-plus, DeepSeek-R1, GPT-3.5, and Llama-70B, while VLMs include Doubao, Qwen-VL, and GPT-4o. This selection provides a broad range of training architectures and accessibility. A full list is provided in Table 2 (Appendix B).

4.2 EVALUATION PROTOCOL

Prompting Strategy. To comprehensively evaluate the performance of different foundation models, we designed three prompting strategies, (1) **PS+I**; (2) **PS+I+DoI**; (3) **PS+DoI**, as illustrated in Figure 3. VLMs were evaluated under all three prompting strategies according to their multi-modal capabilities. In contrast, LLMs were evaluated only under the **PS+DoI** setting, reflecting their text-only input constraints. This design enables a systematic comparison across modalities: (i) whether textualizing diagrams improves reasoning, (ii) whether incorporating schematics enhances performance, and (iii) how visual versus textual representations differentially affect outcomes.

Majority Voting. To mitigate random variability in the model’s output, we adopt a robust evaluation protocol. For each problem and prompt strategy, we generated five independent responses. The final answer was then determined using a majority-vote mechanism, which was implemented using the DeepSeek-V3-0324 (SiliconCloud, 2025) API. This helps to ensure the reliability of our results.

Human Evaluation. The mainstream approach in recent work for evaluation is to use LLMs-as-Judge (Li et al., 2025a). Our pilot study tested on the Qwen-plus (Alibaba Cloud, 2025a) API on 200 sample problems showed that automated grading achieved an 83% agreement with expert judgments. Despite this encouraging result, we ultimately chose to rely on manual evaluation by human experts for the full dataset, given its manageable size and the importance of performing detailed error analyses that automated systems currently cannot provide with sufficient reliability.

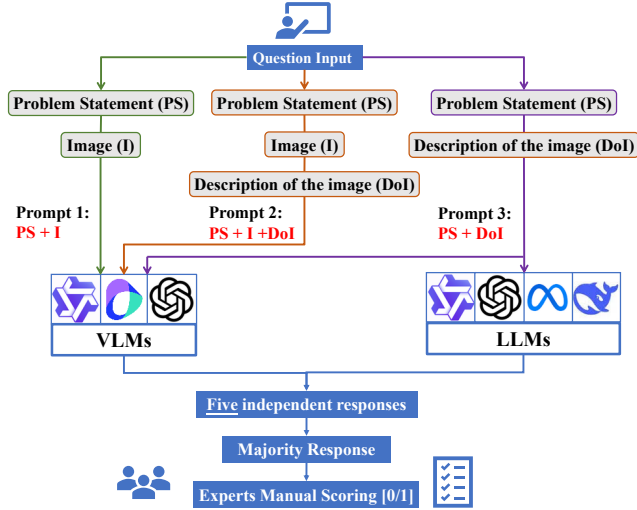


Figure 3: Each problem is tested under 14 model–prompt settings (three strategies \times three VLMs, plus PS+DoI \times five LLMs). For each setting, five responses are generated, majority-voted, and scored by experts with binary labels (1 = correct, 0 = incorrect)

For the human evaluation, the DoI annotation expert team collectively examined each model-generated response across all 1,065 problems. The evaluation proceeded in two stages: first, verifying whether the reasoning process followed a logically valid sequence of steps, and second, checking whether the final answer was correct. A response was awarded a score of 1 only if both criteria were satisfied; otherwise, it received a score of 0. In cases where decisions were difficult or the outcome was ambiguous, the evaluators engaged in discussion until consensus was reached. This process not only provided high-quality ground truth labels but also enabled the identification of systematic error patterns. The overall model performance is reported using Accuracy.

4.3 RESULTS

We report results for all compared foundation models with different prompting strategies for our proposed benchmark dataset SoM-1K in Figure 4. We have the following observations: (1) The best-performing model, Qwen-plus achieved an accuracy of 56.6% using the PS+DoI prompt strat-

egy, while GPT-3.5 scored the lowest at 1.0%. The observed low ratings highlight the significant challenges that current foundation models face when addressing engineering problems. (2) The top-performing models were **Qwen-plus (56.6%)**, **Deepseek-R1 (52.4%)**, and **Doubao (48.5%)**, all of which achieved their best results using the **PS+DoI prompting strategy**. It is also interesting to note that, for VLMs like Doubao and GPT-4o, including an image in the prompt (PS+I or PS+DoI+I) barely improves performance compared to the text-only PS+DoI prompt. (3) With the exception of Doubao, text-only reasoning models (**Qwen-plus**, **DeepSeek-R1**, and **GPT-oss-120b**) generally outperformed the VLMs (**Qwen-VL** and **GPT-4o**) evaluated in this work. (4) Among open-source models, larger LLMs generally perform better: DeepSeek-R1 (**671B**) achieves **52.4%** accuracy, GPT-oss (**120B**) **39.0%**, and Llama (**70B**) only **9.1%**.

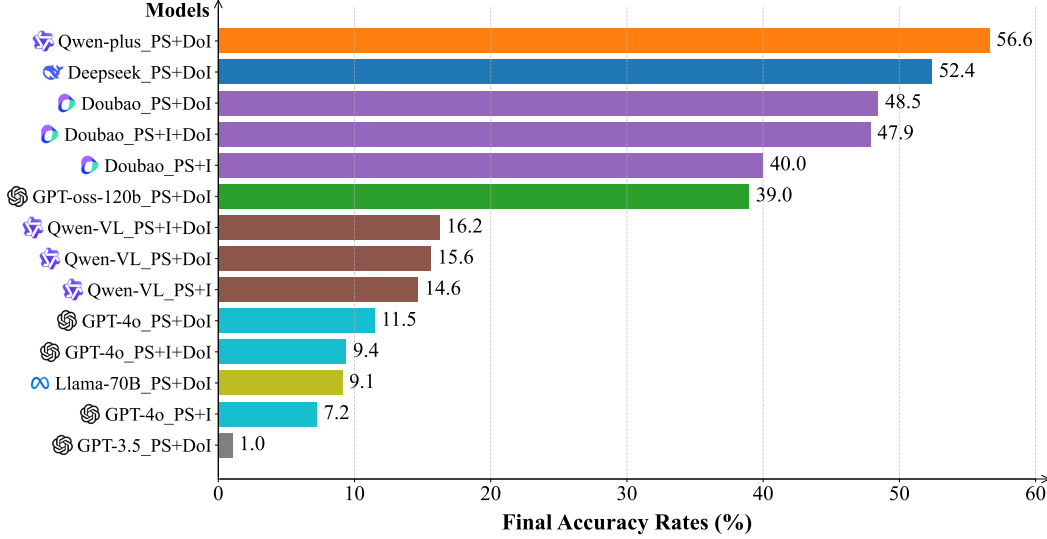


Figure 4: Accuracy for each evaluated model with different prompting strategies.

These findings indicate (1) VLMs’ limited capabilities in interpreting and integrating domain-specific information from schematic diagrams, suggesting the need for further advancement; and (2) the relative effectiveness of well-structured textual information for aiding foundation models’ complex problem-solving, especially for larger models.

5 DISCUSSIONS

5.1 WHAT TYPES OF ERRORS CAN FOUNDATION MODELS MAKE IN SOM-1K?

To gain deeper insights into the capabilities of foundation models in solving mechanics problems, we engaged human experts to examine 100 problems that none of the models were able to solve. Following a thorough manual review, we developed a comprehensive error taxonomy, illustrated with examples in Figure 5. Specifically, we classify the errors into four distinct categories:

- **Type K (Knowledge-based Error):** Failing to apply correct domain knowledge, e.g., misjudging internal loads in a structure.
- **Type C (Calculation Error):** Correct formulas used but numerical results are wrong.
- **Type E (Extraction Error):** Failing to interpret or extract information from the prompts, producing misaligned answers.
- **Type O (Other Error):** Incomplete solutions or responses that entirely miss the problem.

Knowledge-based (K) and calculation (C) errors typically indicate that the model has grasped the problem but faltered in recalling domain knowledge or performing arithmetic mistakes that could potentially be mitigated through external tools, e.g., retrieval augmentation and computational simulations. In contrast, extraction (E) and other (O) errors demonstrate that the model either failed to

interpret the problem correctly or was unable to construct a coherent solution pathway. Such failures point to a breakdown in comprehension and reasoning, directly constraining the model’s ability to engage with mechanics problems.

K (Knowledge-based Error)	C (Calculation Error)
Response: <ul style="list-style-type: none"> Taking moments about point D: $V_B \cdot a - P_2 \cdot b = 0$ 	Response: <ul style="list-style-type: none"> It is calculated that: $d \geq \left(\frac{16000}{30\pi} \right)^{1/3} \approx 2.123 \text{ cm}$
Comment: The sign for P_2 should be positive and the correct equation should be: $V_B \cdot a + P_2 \cdot b = 0$	Comment: The result is miscalculated and the correct answer should be: $d \geq \left(\frac{16}{30000\pi} \right)^{1/3} \approx 55.4 \text{ mm}$
E (Extraction Error)	O (Other Error)
PS: Draw the moment diagrams for the beam using the method of superposition. Response: <ul style="list-style-type: none"> 2. Segment $5 < x \leq 20$ (span A to B) Bending moment formula: $M = -525 + 15s - \frac{s^3}{18}$ Parameter definition: $s = x - 5$ 	Response: <ul style="list-style-type: none"> 1. Primary Structure:[...] 2. Redundant Structure:[...] Feel free to ask if you have any specific questions or need further clarification on this process. End of the response.
Comment: The method of superposition is not used.	Comment: The final solution is not provided.

Figure 5: Description of the four error types.

For each problem among the 100 error cases, the annotation team manually reviewed the majority response from each model and categorized it into one of four predefined error types (K, C, E, O). The reviewers first examined whether the response contained a complete problem-solving process. If the model either failed to provide a full solution or made no attempt to solve the problem, the response was labeled as Type O. Otherwise, the reviewers carefully traced the solution from the beginning, identified the earliest mistake, and assigned it to Type E, C, or K.

5.2 WHAT ERRORS DO DIFFERENT FOUNDATION MODELS MAKE?

The distribution of error types for all evaluated models with different prompting strategies was then computed as the percentage of responses in each category out of the 100 problems. These proportions are reported in Figure 6 as **Percentage**. Because O and E represent the most severe error types, we also report their combined proportion (O+E) to emphasize the overall prevalence of these critical failures.

As shown in Figure 6, while all models failed on these 100 problems, their error distributions differed markedly. In particular, Figure 6(a–c) show that GPT-3.5, GPT-4o_PS+I, Qwen-VL_PS+I, and Llama-70B exhibited a high proportion of Type O and Type E errors, with more than 39% of their failures reflecting a fundamental misunderstanding of the problem. This pattern is consistent with their lower overall accuracy in Figure 4. In contrast, Qwen-plus and DeepSeek-R1 demonstrated substantially fewer critical errors (Type O+E error rates of 7% or less), which aligns with their stronger overall performance. These results suggest that the latter models possess a more reliable grasp of the underlying problem-solving logic.

Notably, under **PS+I**, **Qwen-VL** and **Doubao** show high Type E errors (34% and 19%) versus Type O errors (5% and 1%), reflecting difficulties in extracting visual information. In contrast, **GPT-4o_PS+I**, **GPT-3.5**, and **Llama-70B** exhibit the opposite trend, with Type O errors dominating within the O+E category, indicating challenges in reaching final solutions in this domain due to limited capabilities.

Our earlier results demonstrated that incorporating DoI enhances the performance of VLMs. To better understand the mechanism driving this improvement, we compared Doubao, Qwen-VL, and

GPT-4o under three prompting strategies. The results, summarized in Figure 6b, show that prompting with DoI markedly reduces the frequency of Type E errors. For instance, Doubao’s Type E error rate dropped from 19% under PS+I to 3% under PS+DoI and 5% under PS+I+DoI. Similarly, Qwen-VL’s Type E error rate decreased from 34% (PS+I) to 6% (PS+DoI) and 5% (PS+I+DoI), while GPT-4o’s rate fell from 10% (PS+I) to 2.0% (PS+DoI) and 3.0% (PS+I+DoI). These substantial reductions highlight DoI’s effectiveness in mitigating misinterpretations of visual information, thereby supporting more accurate problem-solving.

As illustrated in Figure 6e, arithmetic errors (Type C) still occur, demonstrating that models may miscompute even when the correct formula is used. Among all four error types, Type K errors are the most frequent (Figure 6f), reflecting gaps in engineering knowledge that could be mitigated via supervised fine-tuning on domain-specific data.

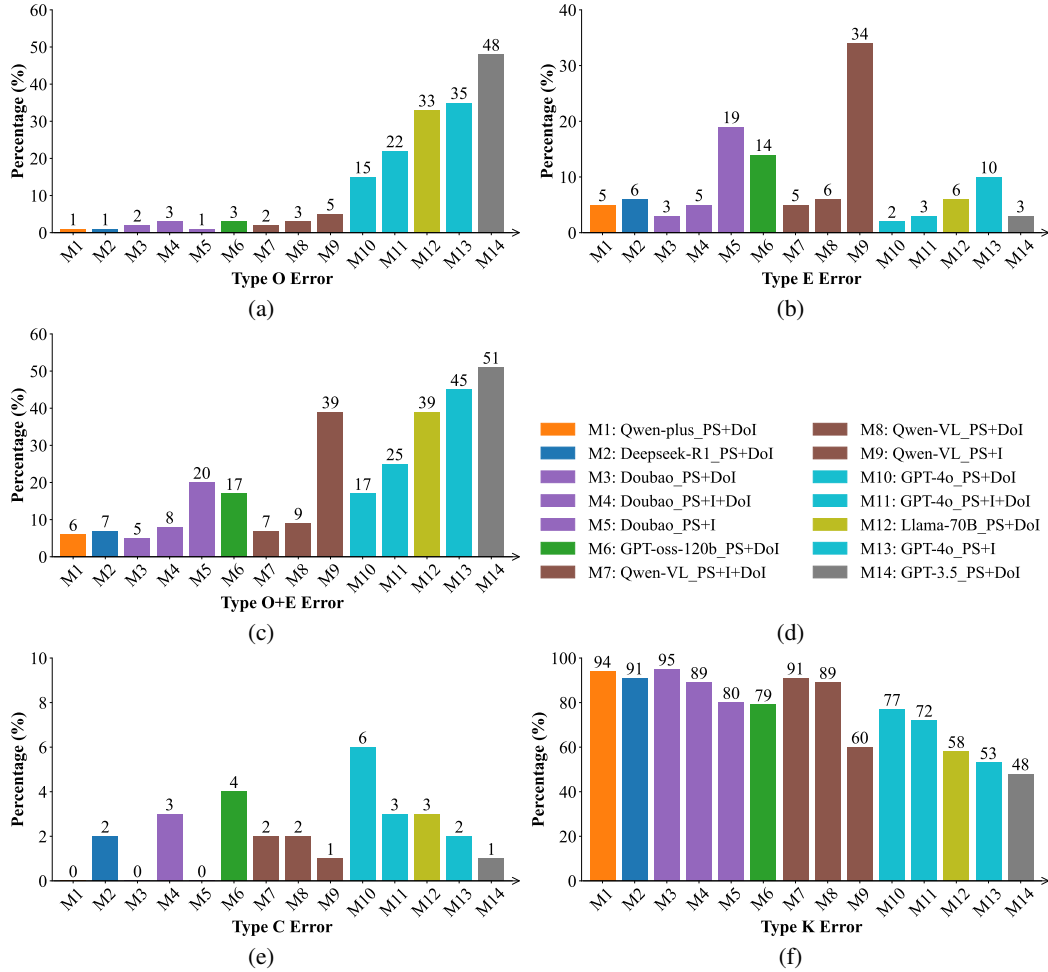


Figure 6: The percentage of error types of each model among 100 questions that all models fail to solve. (a) Type O, (b) Type E, (c) Type O+E, (d) legend, (e) Type C, (f) Type K.

5.3 CAN FOUNDATION MODELS PROVIDE BETTER SOLUTIONS THAN TEXTBOOKS?

While foundation models generally exhibit limited capabilities in solving SoM problems, we find that interestingly, foundation models can sometimes generate better answers. As shown in Figure 8 (Appendix B), the correct solutions generated by Qwen-plus is not only correct but also more detailed and pedagogically structured than the textbook solution. This highlights the potential of foundation models in educational applications, where they can provide richer and more comprehensive explanations for students, particularly for problems requiring multi-step derivations.

6 CONCLUSION

This study introduced SoM-1K, a novel multimodal benchmark for evaluating the problem-solving abilities of foundation models in strength of materials. Unlike previous text-only benchmarks, SoM-1K uses a combination of text and schematic diagrams to provide a more realistic and rigorous evaluation. Our findings reveal that even the most advanced LLMs and VLMs struggle with these complex, domain-specific engineering problems, showing significant limitations in their reasoning capabilities. We also demonstrated that using DoI as a prompting strategy dramatically improves performance by reducing misinterpretation errors, suggesting that for current models, well-structured textual input is a more reliable foundation for complex reasoning than raw visual data.

Future research should focus on expanding the scope of multimodal benchmarks beyond the current limitations of SoM-1K to include more advanced engineering domains like structural dynamics, plasticity, and nonlinear mechanics. The persistent challenges observed in diagram-based reasoning and calculation errors highlight a critical need for future models to enhance their **multimodal reasoning capabilities** and to integrate more effectively with specialized tools. This will enable them not only to solve complex problems but also to reliably generate accurate scientific diagrams, such as internal force diagrams or deformation shapes, which remains a significant hurdle for current foundation models.

ACKNOWLEDGMENTS

This work was supported by the Fundamental Research Funds for the Central Universities (China). The authors are grateful to Prof. Jianhui Luo, College of Civil Engineering, Hunan University, for his valuable feedback and guidance throughout this project.

REFERENCES

- Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. Large language models for mathematical reasoning: Progresses and challenges, 2024. URL <https://arxiv.org/abs/2402.00157>.
- Alibaba Cloud. Qwen-plus-2025-07-28, 2025a. Available at: <https://qwenlm.github.io>.
- Alibaba Cloud. Qwen-vl-max-2025-04-08, 2025b. Available at: <https://qwenlm.github.io>.
- Anton Bakhtin, Laurens van der Maaten, Justin Johnson, Laura Gustafson, and Ross Girshick. Phyre: A new benchmark for physical reasoning, 2019. URL <https://arxiv.org/abs/1908.05656>.
- Arne Bewersdorff, Christian Hartmann, Marie Hornberger, Kathrin Seßler, Maria Bannert, Enkelejda Kasneci, Gjergji Kasneci, Xiaoming Zhai, and Claudia Nerdel. Taking the next step with generative artificial intelligence: The transformative role of multimodal large language models in science education. *Learning and Individual Differences*, 118:102601, February 2025. ISSN 1041-6080. doi: 10.1016/j.lindif.2024.102601. URL <http://dx.doi.org/10.1016/j.lindif.2024.102601>.
- Markus J. Buehler. Melm, a generative pretrained language modeling framework that solves forward and inverse mechanics problems, 2023. URL <https://arxiv.org/abs/2306.17525>.
- ByteDance. Doubao-1.5-thinking-vision-pro-250428, 2025. Available at: <https://www.doubao.com>.
- Chinese Society of Theoretical and Applied Mechanics and Zhou Peiyuan Foundation. Official website of national zhou peiyuan undergraduate mechanics competition. <http://zpy.cstam.org.cn/>, 2025.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam

- Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayanan Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways, 2022. URL <https://arxiv.org/abs/2204.02311>.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL <https://arxiv.org/abs/2110.14168>.
- Hongliang Dai. *Solutions to Mechanics of Materials Exercises: Graduate Entrance Exam Guide*. Hunan University Press, Changsha, China, 7 2015. ISBN 978-7-5667-0909-7.
- DeepSeek. Deepseek-rl-0528, 2025. Available at: <https://www.deepseek.com>.
- Anna C. Doris, Daniele Grandi, Ryan Tomich, Md Ferdous Alam, Mohammadmehdi Ataei, Hyunmin Cheong, and Faez Ahmed. Designqa: A multimodal benchmark for evaluating large language models’ understanding of engineering documentation, 2024. URL <https://arxiv.org/abs/2404.07917>.
- James M. Gere and Barry J. Goodno. *Mechanics of Materials*. Cengage Learning, 7th edition, 2009. ISBN 978-0-534-55397-5.
- Weilin Guo and Dongxing Liu (eds.). *Cailiao Lixue I (5th Edition) Tongbu Fudao ji Xiti Quanjie [Mechanics of Materials I (Fifth Edition) Synchronous Guide and Complete Solutions]*. China Water & Power Press, Beijing, August 2010. ISBN 978-7-5084-7743-5.
- Yunzhuo Hao, Jiawei Gu, Huichen Will Wang, Linjie Li, Zhengyuan Yang, Lijuan Wang, and Yu Cheng. Can mllms reason in multimodality? emma: An enhanced multimodal reasoning benchmark, 2025. URL <https://arxiv.org/abs/2501.05444>.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021. URL <https://arxiv.org/abs/2009.03300>.
- R. C. Hibbeler. *Structural Analysis*. Pearson, 8 edition, 2012. ISBN 978-0132576954.
- Mengsheng Huang. *Solutions to Mechanics of Materials Exercises*. China Electric Power Press, Beijing, China, 2009. ISBN 978-7-5083-9060-4.
- Deepak Kumar and Anil Agrawal. Advancing bridge infrastructure management through artificial intelligence: A comprehensive review. *International Journal of Bridge Engineering, Management and Research*, 2(3):214250021–1:18, Jul. 2025. doi: 10.70465/ber.v2i3.45. URL <https://ijbembr.org/index.php/ber/article/view/45>.
- Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, Yuhuai Wu, Behnam Neyshabur, Guy Gur-Ari, and Vedant Misra. Solving quantitative reasoning problems with language models, 2022. URL <https://arxiv.org/abs/2206.14858>.
- Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, et al. From generation to judgment: Opportunities and challenges of llm-as-a-judge, 2025. In *EMNLP*, 2025a.
- Ming Li, Jike Zhong, Tianle Chen, Yuxiang Lai, and Konstantinos Psounis. Eee-bench: A comprehensive multimodal electrical and electronics engineering benchmark, 2025b. URL <https://arxiv.org/abs/2411.01492>.

- Jiachen Liu, Ziheng Geng, Ran Cao, Lu Cheng, Paolo Bocchini, and Minghui Cheng. A large language model-empowered agent for reliable and robust structural analysis, 2025. URL <https://arxiv.org/abs/2507.02938>.
- Zhihan Liu, Yubo Chai, and Jianfeng Li. Toward automated simulation research workflow through llm prompt engineering design. *Journal of Chemical Information and Modeling*, 65(1):114–124, December 2024. ISSN 1549-960X. doi: 10.1021/acs.jcim.4c01653. URL <http://dx.doi.org/10.1021/acs.jcim.4c01653>.
- Degao Ma. *Mechanics of Materials: Exercises and Detailed Solutions, 5th Edition*. Yanbian University Press, Yanji, China, 7 2011. ISBN 978-7-5634-1786-5.
- Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge, 2019. URL <https://arxiv.org/abs/1906.00067>.
- Meta AI. Llama-3.3-70b-instruct, 2024. Available at: <https://ai.meta.com/llama>.
- Bo Ni and Markus J. Buehler. Mechagents: Large language model multi-agent collaborations can solve mechanics problems, generate new data, and integrate knowledge, 2023. URL <https://arxiv.org/abs/2311.08166>.
- Nripesh Niketan, Arunima Santhoshkumar, and Hadj Batatia. *Integrating External Tools with Large Language Models (LLMs) to Improve Accuracy*, pp. 409–421. Springer Nature Singapore, 2025. ISBN 9789819617586. doi: 10.1007/978-981-96-1758-6_34. URL http://dx.doi.org/10.1007/978-981-96-1758-6_34.
- OpenAI. Gpt-3.5-turbo-0125, 2023. Available at: <https://openai.com>.
- OpenAI. Gpt-4o, 2024. Available at: <https://openai.com>.
- OpenAI. Gpt-oss-120b, 2025. Available at: <https://github.com/openai>.
- Cyril Picard, Jürg Schiffmann, and Faez Ahmed. Dated: Guidelines for creating synthetic datasets for engineering design applications, 2023. URL <https://arxiv.org/abs/2305.09018>.
- Cyril Picard, Kristen M. Edwards, Anna C. Doris, Brandon Man, Giorgio Giannone, Md Ferdous Alam, and Faez Ahmed. From concept to manufacturing: Evaluating vision-language models for engineering design, 2024. URL <https://arxiv.org/abs/2311.12668>.
- Alexander Scarlatos, Naiming Liu, Jaewook Lee, Richard Baraniuk, and Andrew Lan. *Training LLM-Based Tutors to Improve Student Learning Outcomes in Dialogues*, pp. 251–266. Springer Nature Switzerland, 2025. ISBN 9783031984143. doi: 10.1007/978-3-031-98414-3_18. URL http://dx.doi.org/10.1007/978-3-031-98414-3_18.
- Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022. URL <https://arxiv.org/abs/2210.08402>.
- Kathrin Seßler, Yao Rong, Emek Gözlüklü, and Enkelejda Kasneci. Benchmarking large language models for math reasoning tasks, 2024. URL <https://arxiv.org/abs/2408.10839>.
- Team SiliconCloud. Deepseek-v3-0324 api. <https://cloud.siliconflow.cn/>, 2025. Accessed via SiliconFlow cloud platform.
- Xunfang Sun, Xiaoshu Fang, and Laitai Guan. *Materials Mechanics I*. Higher Education Press, Beijing, China, 5 edition, 7 2009. ISBN 978-7-04-026473-9.
- Chuan Tian and Yilei Zhang. Optimizing collaboration of llm based agents for finite element analysis, 2024. URL <https://arxiv.org/abs/2408.13406>.

- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding, 2019. URL <https://arxiv.org/abs/1804.07461>.
- Lintao Wang, Encheng Su, Jiaqi Liu, Pengze Li, Peng Xia, Jiabei Xiao, Wenlong Zhang, Xinnan Dai, Xi Chen, Yuan Meng, Mingyu Ding, Lei Bai, Wanli Ouyang, Shixiang Tang, Aoran Wang, and Xinzhu Ma. Physunibench: An undergraduate-level physics reasoning benchmark for multimodal models, 2025. URL <https://arxiv.org/abs/2506.17667>.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL <https://arxiv.org/abs/2201.11903>.
- Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. Clevrer: Collision events for video representation and reasoning, 2020. URL <https://arxiv.org/abs/1910.01442>.

APPENDIX

A DATASET EXAMPLES

To illustrate the diversity of SoM-1K, Figure 7 shows one representative problem from each dataset category.

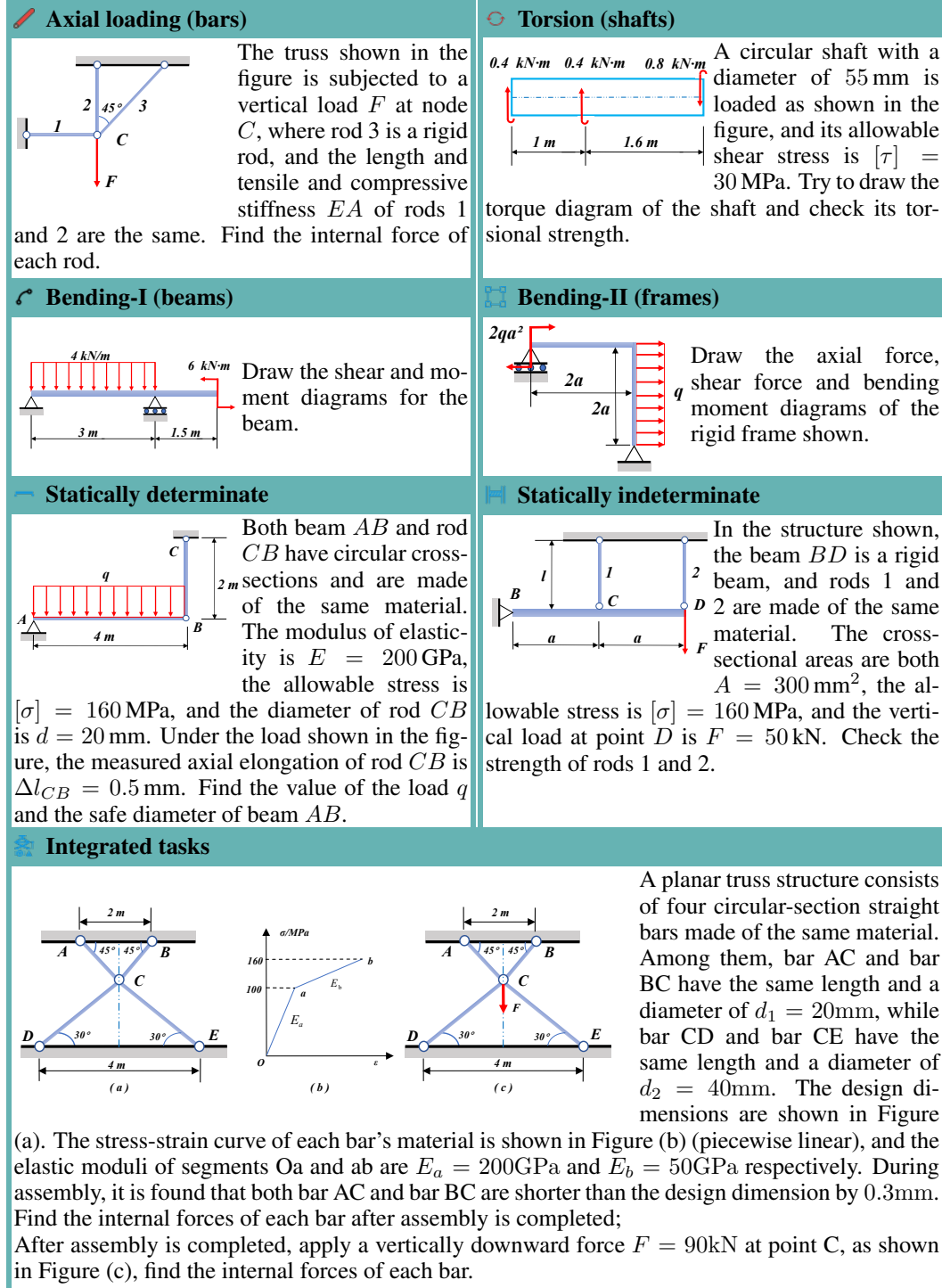


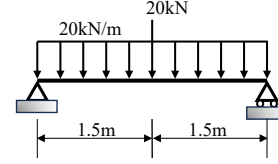
Figure 7: Illustrative examples of one representative problem from each category in the SoM-1K dataset.

B ADDITIONAL TABLES AND FIGURES

Problem statement (PS)

A simply supported beam with a circular tube cross-section is loaded as shown in the figure. It is known that $d/D = 7/10$, and the allowable stress of the material $[\sigma] = 160 \text{ MPa}$. Try to find the required outer diameter D and inner diameter d .

Image (I)



Ground Truth (GT)

The maximum bending moment of the circular tube section:

$$M_{\max} = \frac{pl}{4} + \frac{ql^2}{8} = 37.5 \text{ kN}\cdot\text{m}$$

From the maximum stress $\sigma_{\max} = \frac{M_{\max}}{W} \leq [\sigma]$, we get:

$$D^3 = \frac{32M_{\max}}{\pi(1 - \alpha^4)[\sigma]}$$

Solving these equations, we obtain:

$$D = 147 \text{ mm}, \quad d = 103 \text{ mm}$$

Qwen-plus Response (better than GT)

1. Force Analysis of the Beam

- Support Reactions

Due to the symmetric load, the support reactions are $R_A = R_B = \frac{80}{2} = 40 \text{ kN}$.

- Maximum Bending Moment

The maximum bending moment occurs at the midpoint of the beam, caused by the combined action of the uniformly distributed load and the concentrated force.

2. Bending Stress Condition

- Bending Stress Formula

$$\sigma_{\max} = \frac{M_{\max}}{W_z} \leq [\sigma]$$

Here, the allowable stress $[\sigma] = 160 \text{ MPa} = 160 \text{ N/mm}^2$. Therefore:

$$W_z \geq \frac{M_{\max}}{[\sigma]} = \frac{37.5 \times 10^6}{160} = 234,375 \text{ mm}^3$$

3. Calculation of Section Modulus for Hollow Circular Section

- Section Modulus for Bending

$$W_z = \frac{\pi(D^4 - d^4)}{32D}$$

Given $\frac{d}{D} = \frac{7}{10} \Rightarrow d = 0.7D$, substituting this into the formula gives:

$$W_z = \frac{\pi D^4(1 - 0.7^4)}{32D} = \frac{\pi D^3 \times 0.7599}{32} \approx 0.0746 D^3$$

- Solving for D

Let $0.0746 D^3 \geq 234,375$, then:

$$D^3 \geq \frac{234,375}{0.0746} \approx 3,141,593 \Rightarrow D \geq \sqrt[3]{3,141,593} \approx 146.5 \text{ mm}$$

- Solving for d

$$d = 0.7D = 0.7 \times 146.5 \approx 102.55 \text{ mm}$$

Figure 8: Comparison between solutions generated by foundation models and textbook ground truth.

Table 2: Overview of evaluated foundation models, including source availability, release year, modality and size.

Model	Full Name	Open Source	Team	Year	Modality	Size
Doubao (ByteDance, 2025)	Doubao-1.5-thinking-vision-pro-250428	Closed	ByteDance	2025	VLM	N/A
Qwen-plus (Alibaba Cloud, 2025a)	Qwen-plus-2025-07-28	Closed	Alibaba Cloud	2025	LLM	N/A
Qwen-VL (Alibaba Cloud, 2025b)	Qwen-VL-Max-2025-04-08	Closed	Alibaba Cloud	2025	VLM	N/A
Deepseek-R1 (DeepSeek, 2025)	Deepseek-R1-0528	Open	DeepSeek	2025	LLM	671B
GPT-oss-120b (OpenAI, 2025)	GPT-oss-120b	Open	OpenAI	2025	LLM	120B
GPT-4o (OpenAI, 2024)	GPT-4o-2024-08-06	Closed	OpenAI	2024	VLM	N/A
GPT-3.5 (OpenAI, 2023)	GPT-3.5-turbo-0125	Closed	OpenAI	2023	LLM	N/A
Llama-70B (Meta AI, 2024)	Llama-3.3-70B-instruct	Open	Meta AI	2024	LLM	70B