

Evaluation of Small Vision-Language Models on Qualitative Mechanical Problems

Henry Fordjour Ansah¹, Shreya Banerjee¹, Pranish Ghimire¹

¹University of New Orleans

{hfansah, pghimir3}@uno.edu, shreyabbanerjee@gmail.com

Abstract

Qualitative mechanical problem-solving (QMPS) refers to solving qualitative problems from the mechanical domain. Qualitative problems can be solved with minimal discipline-specific information, without any robust quantitative calculation, generally by using qualitative reasoning and commonsense knowledge. QMPS is a vital aspect of human intelligence that allows us to tackle a wide range of tasks, from simple everyday ones such as turning on a tap to complex tasks in highly demanding and well-paying jobs in various fields, e.g., emergency medicine, plumbing, driving, etc. Employers often use the *Bennett Mechanical Comprehension Test (BMCT)* to evaluate job candidates' ability to solve such problems. In this work, we assess two state-of-the-art multimodal models, Gemma-3 [Team *et al.*, 2025] (*Huggingface ID: google/gemma-3-4b-it*) and Qwen-VL [Bai *et al.*, 2023] (*Huggingface ID: Qwen/Qwen2.5-VL-7B-Instruct*), on their ability to interpret mechanical problem images by eliciting a step-by-step *chain of thought (CoT)* and a final answer. Each image inherently encodes ground-truth qualitative facts, such as contact points in gears, support relations, and relative weights, which we use to evaluate each model's spatial and commonsense reasoning capabilities. We assess each chain for coherence, completeness, and logical progression to assess each model's thought process, and final answers are compared to verified solutions to measure accuracy.

1 Introduction

Qualitative problems differ from quantitative problems in that the former can be solved with scant discipline-specific information and without any robust quantitative calculation.

- **Quantitative Problem Solving:** Solving a problem based on the actual quantity of attributes (value of

length, volume, distance, etc.) as employed in the particular sciences (e.g. physics).

- **Qualitative Problem Solving:** Solving a problem based on the quality of attributes (comparison of size, appearance, symmetry, volume, length, etc.) as employed in everyday cognition.

Qualitative mechanical problem-solving (QMPS) is a method of solving mechanical problems through *qualitative reasoning (QR)* [Forbus, 1996], [Weld and de Kleer, 2013] rather than through robust quantitative calculations requiring equations, theorems, formulae, etc., from mechanics/physics textbooks. This concept is brilliantly captured in Hayes work [Hayes, 1978; Hayes, 1985] where qualitative reasoning about fluid behavior is leveraged from human general knowledge rather than physics proper. QMPS is integral to human intelligence, allowing individuals to perform everyday tasks (such as hanging a picture frame on a wall) as well as complex professional tasks that pay well in today's economy (e.g., in emergency medicine, plumbing, and the use of hydraulic machinery), using commonsense knowledge without the need for discipline-specific information or rigorous calculations. This skill is lacking in current artificial intelligence (AI) and robotics [Banerjee, 2023], marking a significant gap in achieving artificial-general intelligence (AGI) in machines.

In this work, we conduct a detailed evaluation of two small vision-language models, Qwen-VL and Gemma-3 on a single A100-40GB GPU, focusing on their ability to solve qualitative mechanical problems. This assessment moves beyond simple answer accuracy by also critically examining the models' underlying thought processes (chain-of-thought). By analyzing both the correctness of their solutions and the coherence and validity of their reasoning, we aim to provide a nuanced understanding of their performance, identify common error patterns, and highlight their respective strengths and weaknesses in this challenging domain.

The following section further describes QMPS, and our motivation for this work. Next, we illustrate our approach and subsequently analyze the results with detailed case studies, followed by discussion and future work.

2 Motivation and Background

The motivation for this work stems from the increasing need to understand and benchmark the reasoning capabilities of small vision-language models (SVLMs), such as Gemma and Qwen-VL, particularly for qualitative problem-solving in specialized domains. While large language models such as GPT-4 [Achiam *et al.*, 2023] have demonstrated impressive feats in several domains, smaller, more efficient models are becoming crucial for broader accessibility and deployment in real-time environments. Evaluating these SVLMs on tasks that require not just language understanding or visual perception in isolation, but the integrated reasoning about mechanical concepts presented visually and described textually, is vital for agents to solve QMPS challenges in the real world at the level of human intelligence, and perhaps beyond.

Under the rubric of *Psychometric AI* [Bringsjord and Schimanski, 2003] that rigorizes the goals of AI to corresponding human-level tests of cognitive power, for QMPS by AI, we focus on problems from the *Bennett Mechanical Comprehension Tests (BMCT-II)* [Bennett, 2008], which we call *BMCT* or *Bennett*, in tandem. The BMCT assesses a human’s ability to understand mechanical concepts and apply them for problem-solving. As such, it has long been used in the real world by many companies, across several different industries, as a recruitment/training tool to evaluate job candidates for roles that require innate mechanical acumen. BMCT problems are multiple-choice questions (MCQ) consisting of one or more diagrams with labels, markings and textual information, a separate or combined story and question (= stem), with a number of options ranging from 3-5. Challenges of solving such problems include representing visual-conceptual relationships, spatial attributes, labels and annotations from the diagrams.

The specific focus on BMCT/Bennett problems was first followed by [Klenk *et al.*, 2011], who reported an analogical approach, whereas recently [Banerjee *et al.*, 2022] showed a formal logicist approach to the Bennett or BMCT problems. We have evaluated the strengths and weaknesses of a purely statistical approach via vision-language models.

3 Methodology

3.1 Dataset or Test Cases

From iPREP BMCT series [iPREP, 2022], we curated 30 benchmark problems from across four scenario types: hydraulics (7), Gears and Belt Drives (6), Acoustics (6) and Heat and Radiation (6) as used in the Bennett test, with each domain having 2 easy-level questions, 2 medium-level questions and 2 hard-level questions, except the hydraulics domain, which has 3 medium-level questions. Each instance of a problem is a graphical depiction of a type of qualitative mechanical problem that encodes all necessary visual cues, such as fluid levels, force vectors, or gear configurations, accompanied by a question and a fixed set of answer choices. We embed each instance into a template with a system prompt

shown below, to elicit a chain-of-thought reasoning from the models in a suitable format.

SYSTEM PROMPT: Write a response that appropriately completes the request. Before answering, think carefully about the question and create a step-by-step chain of thoughts to ensure a logical and accurate response. Put the step-by-step chain of thoughts in an think tag (<think>...</think>) and put your final answer which should be from the options provided only in an answer tag (<answer>...</answer>).

3.2 Categories of Failure and Success Cases

In order to understand the sources of errors and unexpected successes in our multimodal reasoning pipeline, we analyze four prototypical cases across the 30 BMCT benchmark problems.

Case 1: Flawed Reasoning Leading to Wrong Answer

In this scenario, the model’s chain of thought (CoT) explicitly mentions all relevant visual facts (high coverage) and the final answer is incorrect. Investigation reveals that the error arises from one or more of the following:

- *Incorrect Assumptions:* The model infers properties not supported by the image (e.g., assuming unequal lever arms when both arms are equal length).
- *Misapplication of Physical Concepts:* Fundamental qualitative laws—such as force equilibrium or torque balance—are applied incorrectly (for example, adding forces when the correct operation is to compare magnitudes qualitatively).
- *Logical Fallacies:* Steps in the CoT contain internal contradictions or invalid deductions (e.g., inferring that a large object must be less dense than a small one because it floats without considering the average density relative to the fluid).

We catalog these error types and compute their frequencies. For each error instance, we also record the CoT quality score to examine whether low coherence or logical progression correlates with flawed concept application.

Case 2: Omission of Crucial Visual Information - Poor Spatial Reasoning

Here, the model’s final answer is incorrect primarily because the CoT fails to mention one or more essential facts encoded in the image. Typical omissions include:

- *Missing Contact Points:* The model does not register which object is in contact with which support or surface, such as in the case of multiple gear connections.
- *Ignored Force Vectors or Pressure Levels:* In hydraulics problems, differential fluid heights are overlooked; in gear problems, gear tooth counts or meshing relationships are omitted.
- *Overlooked Geometric Relations:* Relative distances or angles (e.g., incline angle, lever arm lengths) are not referenced.

Case 3: Correct Answer Despite Flawed Reasoning

Occasionally, the model arrives at the correct final answer even though the CoT exhibits clear reasoning flaws. These cases fall into two subtypes:

- *Lucky Guessing*: The model hallucinates intermediate facts or applies concepts incorrectly but somehow incorrectly arrives at the correct answer by chance.
- *Partial Reasoning Suffices*: Despite an incomplete or partially incorrect chain, the model captures a minimal set of essential facts correctly, enabling a correct qualitative judgment.

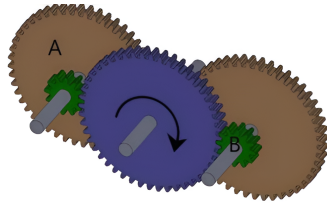
Case 4: Correct Answer with Flawless Reasoning

We will also analyze a case where both models not only produce the correct answer but also display coherent, logically sound chains of thought.

3.3 Results and Analysis

Table 1 provides an overview of both models’ performance across all 30 problems, reporting their overall accuracy as well as the percentage of flawless versus flawed reasoning and failures due to poor spatial reasoning. Below, we further delve into each category with a demonstration, exploring how reasoning quality and spatial understanding contribute to overall success or failure.

Case 1: Flawed Reasoning Leading to Wrong Answer



The middle gear rotates the smaller gears on both sides. If gear A is the larger gear on the left and gear B is the smaller gear on the right, which gear will rotate faster?
(If equal, mark C.)

- ☐ A.
☐ B.
☐ C.

Figure 1: Sample qualitative mechanical reasoning problem from BMCT in the gear domain

The figure 9 shows a peculiar case where both models provide an incorrect answer due to flawed reasoning. In this problem, a central driving gear (purple) meshes with two smaller gears: one labeled B (right) and the other rigidly fixed to a larger gear labeled A (left). Both meshes occur at the same tangential (rim) speed. The question asks: given that the middle gear drives both sides equally, which gear, A or B, will end up rotating faster?

Correct Reasoning Trace

As we can see, the middle gear directly meshes with the smaller gears on both sides. We know from the basic physics principle that the tangential speed of meshed

gears is identical. Since both smaller gears mesh with the middle gear, both would have the same tangential speed. Both are of the same size (radius), hence angular speed is same too.¹ Consequently, both smaller gears rotate their shafts at identical rates. Since each outer wheel (A and B) is rigidly fixed to its respective pinion, they too turn at the same angular velocity. Hence, gear A and gear B rotate at the same speed.

Both Gemma and Qwen-VL arrive at the same (incorrect) conclusion: “Gear B rotates faster than Gear A”, and their chains of thought portray distinct logical fallacies when compared to the correct reasoning.

Case 1 - CoT Illustration - Gemma

The question describes a gear system where a middle gear rotates smaller gears on both sides. Gear A is larger on the left and Gear B is smaller on the right.

We need to determine which gear will rotate faster.

The relationship between the speed of a driving gear and a driven gear in a gear system is inversely proportional: if the driving gear rotates faster, the driven gear rotates slower, and vice versa.

The number of teeth on the gears dictates this speed ratio. A larger gear (with more teeth) will cause the meshed smaller gear to spin more times per single rotation of the larger gear.

Because Gear A is larger than Gear B, when Gear A rotates it drives Gear B more quickly (since B must complete more rotations to match the same rim distance). Therefore, Gear B will rotate faster than Gear A.

Figure 2: Gemma’s (incorrect) CoT for gear-speed problem.

Gemma’s reasoning begins correctly by noting that the middle gear meshes with both side gears and that this mesh enforces identical tangential velocities at each contact point (highlighted in green). Nevertheless, it subsequently invokes the standard inverse proportionality between driving-gear and driven-gear speeds, and the associated teeth-count relationship, as if Gear A directly drives Gear B (highlighted in red). This rule, however, is inapplicable to the present configuration, in which both side gears are independently driven by the central gear. By misapplying this driver–driven proportionality and treating the larger left wheel (A) as the “driver” with more teeth, Gemma deduces that the smaller right wheel (B) must rotate more rapidly.

¹For rigorous treatment, we might recall the fundamental relation:

$$v = r \cdot \omega,$$

where v is the tangential (rim) speed, r is the gear’s radius, and ω is its angular (rotational) speed. Because both the tangential speed v and the radius r are the same for the two meshed gears, their angular speeds ω must also be the same.

Table 1: Summary of Performance of Qwen-VL and Gemma on 30 BCMT Problems

	Accurate		Inaccurate	
Gemma	36.67%		63.33%	
	Correct Reasoning 30%	Incorrect Reasoning 6.67%	Flawed Reasoning 40.00%	Missed Visual Cue 23.33%
Qwen	46.67%		53.33%	
	Correct Reasoning 40.00%	Incorrect Reasoning 6.67%	Flawed Reasoning 33.33%	Missed Visual Cue 20.00%

Case 1 - CoT Illustration - Qwen-VL

To determine which gear will rotate faster, we need to consider the relationship between the gears and their sizes. Here’s the step-by-step reasoning:

1. Identify the gears: Gear A is the larger gear on the left, and Gear B is the smaller gear on the right.

2. Understand the rotation: The middle gear rotates both smaller gears. Since the middle gear is the same size for both smaller gears, it means that the speed at which the middle gear rotates will be the same for both smaller gears.

3. Determine the speed: Because the middle gear is the same size for both smaller gears, the smaller gear on the right (Gear B) will rotate faster than the smaller gear on the left because it has fewer teeth and thus a higher rotational speed for the same linear distance traveled by the middle gear.

Given this analysis, Gear B will rotate faster.

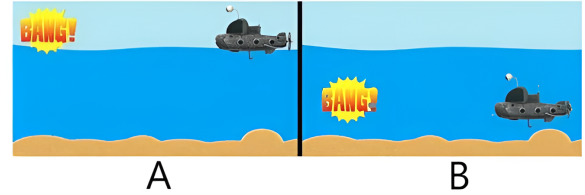
Figure 3: Qwen-VL’s (incorrect) CoT for gear-speed problem.

Qwen-VL correctly observes that the central gear meshes with two identical smaller gears, thereby imposing the same tangential velocity at each interface. Nevertheless, it then inexplicably asserts that the right-hand gear (B) rotates more rapidly because “it has fewer teeth and thus a higher rotational speed than the smaller left gear.” This reasoning is doubly flawed: the two driven gears seem identical in tooth count, and a gear’s angular velocity under a fixed tangential input depends solely on its own radius rather than on any comparison of tooth counts between them. By conflating tooth count with rotational speed, Qwen-VL introduces an unfounded assumption that leads to the incorrect conclusion that gear B spins faster than gear A.

Flaws in the chains of thought produced by both Gemma and Qwen directly lead to their erroneous conclusion that the smaller right-hand gear spins faster. In each case, the models hallucinate dependencies that simply do not exist in this configuration: Gemma treats Gear A as the “driver” and reverses the teeth-speed relationship, while Qwen invents a teeth-count disparity between identical driven gears. These hallucinations of

nonexistent facts and misused physical laws compromise the integrity of every subsequent inference, ultimately yielding an inaccurate final answer, pointing towards the crucial need for verifiable proof for their answers.

Case 2: Omission of Crucial Visual Information - Poor Spatial Reasoning



In which scenario will the submarine hear the blast first?

(If at equal time, mark C.)

- ☐ A.
☐ B.
☐ C.

Figure 4: Sample qualitative mechanical reasoning problem from BMCT in the acoustics domain

This problem illustrates a case in which both models fail to answer correctly because they focus on the wrong visual cues. It presents two scenarios for a submarine detecting a blast: in Scenario A the explosion occurs in air and the sound must travel entirely through air before reaching the submarine which is on the surface of the water, whereas in Scenario B the blast originates in the water and propagates through water directly to the submarine which is also underwater. The task is to decide which scenario yields the earliest detection of the sound.

Correct Reasoning Trace

Sound is a mechanical wave that requires a medium for propagation, and its speed depends on the medium’s properties². In this problem, the submarine must detect the same blast through either air or water. Because water has a much higher density (and bulk modulus) than

²This is backed by the relation,

$$v = \sqrt{\frac{B}{\rho}},$$

where B is the bulk modulus and ρ the density.

air, sound travels substantially faster in water, approximately 1,482 m/s versus 343 m/s in air (and even 5,941 m/s in steel). Consequently, when the explosion occurs in the water (Scenario B), the acoustic wave propagates through the denser medium and reaches the submarine more quickly than in Scenario A, where the sound must traverse the slower, less dense air before reaching the submarine. Therefore, the submarine will hear the blast first in Scenario B.

Case 2 - CoT Illustration - Gemma

The question asks in which scenario the submarine will hear the blast first. We need to analyze the two scenarios depicted.

Scenario A shows the submarine moving towards a blast that's already happened. The blast occurs first and the submarine is travelling towards it.

Scenario B shows the submarine moving away from a blast that's already happened. The blast occurs first and the submarine is moving away from it.

Since the question asks which scenario the submarine will hear the blast first, we need to consider the time it takes for the sound to travel. Sound travels at a constant speed. The submarine will hear the blast as it approaches it. In scenario A, the submarine is moving towards the blast, so it will hear it sooner. In scenario B, it's moving away from the blast, so it will hear it later.

Therefore, the answer must be A.

Figure 5: Gemma's (incorrect) CoT for the acoustic problem.

Gemma's CoT begins by misreading the static illustrations as depicting a moving submarine: it interprets "Scenario A" as the vessel racing toward an already occurred blast and "Scenario B" as it fleeing from one, even though there is no such indication that the submarine is moving. By inventing this nonexistent motion cue, Gemma then invokes irrelevant kinematic reasoning, comparing approach versus recession times rather than attending to the true medium-dependent propagation speeds. This misidentification of the visual scenarios leads it to apply an inappropriate "sound travels at constant speed relative to a moving observer" argument and thus to the wrong conclusion that the submarine hears the blast sooner in A.

Qwen-VL similarly misreads the illustrations by presuming an apparent disparity in separation between the submarine and the blast in the two scenarios. It treats Scenario A as if the vessel lies closer to the detonation and Scenario B as if it lies farther away, even though both images depict identical horizontal distances. By inventing this spurious "distance" cue, Qwen then invokes an irrelevant causal factor, emphasizing proximity, rather

than recognizing that the acoustic medium is the true determinant. This leads it to conclude, without justification, that the closer submarine in A will hear the blast first, thereby overlooking the critical medium-dependent speed difference and arriving at a false solution.

Case 2 - CoT Illustration - Qwen-VL

The question asks in which scenario the submarine will hear the blast first.

The key factor here is the distance between the submarine and the explosion.

- In Scenario A, the submarine is closer to the explosion compared to Scenario B. - Sound travels faster through water than through air, but the primary consideration here is the physical distance.

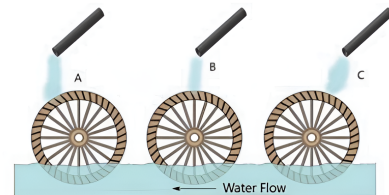
Since the submarine is closer to the explosion in Scenario A, it will hear the blast first.

Final Answer: A.

Figure 6: Qwen-VL's (incorrect) CoT for the acoustic problem.

By eliciting explicit chains of thought from these multimodal models, we gain some level of insight into which visual features they attend to and how they integrate those cues into their qualitative reasoning. Instead of treating the models as black boxes whose final response merely reflects internal knowledge, the step-by-step rationales reveal the precise assumptions they derive from each image.

Case 3: Correct Answer Despite Flawed Reasoning



Three turbines A, B, and C are installed in a river with water flowing to rotate them. Water also falls from pipes on their blades to provide extra speed. Judging by the location of pipes, which turbine will run the fastest?

- ☐ A.
- ☐ B.
- ☐ C.

Figure 7: Sample qualitative mechanical reasoning problem from BMCT in the hydraulics domain

In this unusual case where the Qwen model luckily arrives at a correct answer, the problem presents a three-turbine setup, each wheel receives the same water current but also a jet of water from a pipe positioned at different points along its blades. Turbine A's pipe feeds water at the left side of its blades, Turbine B's pipe is directly over its pivot, and Turbine C's pipe delivers water at the right side of its pivot. Although a rigorous torque analysis would show that only Turbine C gains a reinforcing

moment from the jet, Qwen instead simplifies the problem to the idea that the speed of the jet water is dependent on where it falls on the blades and that is what determines which blade rotates faster. By focusing solely on water velocity rather than the resultant moment, it nevertheless, and purely by chance, identifies Turbine C (correct answer) as the fastest, despite wrong assumptions.

Case 3 - CoT Illustration - Qwen

To determine which turbine will run the fastest, we need to consider the amount of water flow and the additional speed provided by the pipes on each blade.

1. Turbine A: The pipe is located at the top of the blade, which means it will provide the least additional speed because the water will have less time to accelerate before hitting the blade.

2. Turbine B: The pipe is located in the middle of the blade, providing more additional speed than Turbine A but less than Turbine C.

3. Turbine C: The pipe is located at the bottom of the blade, which means it will provide the most additional speed because the water will have more time to accelerate before hitting the blade.

Given this information, Turbine C will have the highest additional speed due to the position of its pipe, making it the fastest.

Figure 8: Qwen-VL’s (incorrect) CoT for the acoustic problem.

Correct Reasoning Trace

In the three-turbine arrangement, each wheel experiences a constant clockwise moment from the river flow, but the falling-water jet’s contribution depends critically on its position relative to the pivot. For Turbine A, the jet strikes ahead of the pivot, producing a counterclockwise moment that opposes the river’s torque. For Turbine B, the jet impacts directly above the pivot and thus generates no additional moment. In contrast, Turbine C’s jet falls behind the pivot, adding a clockwise moment that reinforces the river-driven torque. Summing these effects shows that only Turbine C gains extra driving moment, making it the fastest runner.

Such “lucky-hit” cases where a model’s flawed internal logic nonetheless produces a correct answer highlight both the brittleness and unpredictability of language models in qualitative reasoning.

Case 4: Accurate Answer and Reasoning

In a fourth evaluation scenario, we examine a problem in which both Gemma and Qwen-VL not only produce the correct answer but also generate coherent, logically sound chains of thought. Here, the cues are unambiguous and the underlying geometric principles are straightforward to apply. Both models accurately extract the necessary details, construct step-by-step reasoning

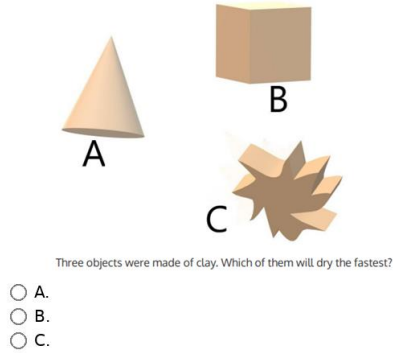


Figure 9: Sample qualitative mechanical reasoning problem from BMCT in the hydraulics domain

that mirrors the ideal solution, and converge on the correct conclusion.

Case 4 - CoT Illustration - Gemma

The question asks which of the three objects made of clay will dry the fastest. Clay dries when water evaporates from its surface. The surface area exposed to the air influences the rate of evaporation. Objects with larger surface areas exposed to air will dry faster.

Let’s analyze each object:

* A (Cone): The cone has a relatively small surface area compared to its volume. A large portion of the clay is inside the cone and not exposed to the air.

* B (Cube): The cube has six faces. Each face has a relatively large surface area.

* C (Fragment): The fragment has a very large surface area due to its many edges and corners.

Comparing the three, the fragment (C) has the largest surface area exposed to the air. Therefore, it will dry the fastest.

Figure 10: Gemma’s (correct) CoT for heat-radiation problem.

In the handful of problems where both Gemma and Qwen-VL delivered flawless chains of thought and correct answers, the visual clues were unambiguous/easy-to-understand and the governing principles were nearly straightforward. In these “easy” cases, the models could reliably map pixels to facts and apply familiar rules without missteps, suggesting that when spatial layouts and qualitative laws are both overt and straightforward, even compact vision-language systems perform excellently. However, this success under ideal conditions also underscores their limitations: as soon as the required visual inference becomes subtle or the physics more nuanced, their spatial reasoning can falter and their qualitative deductions go astray.

Case 4 - CoT Illustration - Qwen-VL

The question asks which of the three clay objects will dry the fastest. The drying rate of clay objects is influenced by their surface area-to-volume ratio. Objects with a higher surface area-to-volume ratio will dry faster because more water can evaporate from their surfaces.

- Object A (cone) has a relatively small surface area compared to its volume.

- Object B (cube) has a moderate surface area-to-volume ratio.

- Object C (sculpture) has a very high surface area-to-volume ratio due to its many protruding edges and corners.

Given this information, the object with the highest surface area-to-volume ratio will dry the fastest.

Figure 11: Gemma’s (correct) CoT for heat-radiation problem.

4 Discussion, Conclusion, and Future Work

The evaluation of Gemma and Qwen on visual reasoning tasks reveals distinct performance profiles and common challenges faced by current language models. While both models demonstrate capabilities in achieving accurate answers with flawless reasoning to a certain extent (Gemma at 30% and Qwen at 40%), a considerable number of errors highlight persistent issues in their reasoning.

A significant portion of errors for both models stems from either flawed reasoning or poor spatial reasoning. The category of “Accurate but Flawed Reasoning,” while resulting in a correct final answer, is particularly noteworthy. It emphasizes that answer-level accuracy alone is insufficient for evaluating true understanding and reliability. For real-life purposes, this points towards the importance of underlying justification for LLM’s answers, in the form of verifiable proof/argument. Flawed reasoning, particularly when it involves the misapplication or hallucination of domain-specific knowledge, points to limitations in the model’s grounding and factual recall when applied to complex problems. Poor spatial reasoning highlights the ongoing challenge of robust visual understanding, which is foundational for tasks requiring interpretation of diagrams, scenes, or visual data. We discuss below how each model fared in all 4 cases below.

Accurate and Flawless Reasoning: Gemma achieved accurate and flawless reasoning in 9 out of 30 cases (30%), correctly identifying key visual elements, applying appropriate principles, and following a coherent logical process to arrive at the right answer. Qwen outperformed Gemma slightly in this category, demonstrating accurate and flawless reasoning in 12 out of 30 cases (40%). In these instances, Qwen reliably

interpreted the visual information and applied sound, step-by-step logic without misapplying any physical laws.

Accurate but Flawed Reasoning: Both models produced correct final answers despite flaws in their chains of thought in 2 out of 30 cases each (6.67%). Gemma’s explanations occasionally contained inaccuracies, such as misstatements of physical principles or minor logical missteps, yet arrived at the right answer by coincidence - a dangerous trend per se. Qwen exhibited a similar pattern - partially logical reasoning or subtle misapplications did not prevent it from selecting the correct option.

Inaccurate Due to Flawed Reasoning: Gemma failed due to flawed reasoning in 12 out of 30 cases (40%). These errors stemmed from incorrect assumptions, hallucinated concepts, or invalid deductive steps leading to an incorrect conclusion. Qwen exhibited this failure mode in 10 out of 30 cases (33.33%).

Inaccurate due to Poor Spatial Reasoning: Gemma’s inability to interpret critical visual details led to errors in 7 out of 30 cases (23.33%). Even when its internal reasoning was logically coherent, misperceiving spatial relationships or missing key image features caused incorrect premises and, therefore, incorrect outcomes. Qwen encountered this issue in 6 out of 30 cases (20%).

We observe that addressing these limitations in small vision-language models requires a multi-pronged approach focusing on enhancing both perceptual and reasoning capabilities. In the context of pure statistical approach for qualitative mechanical reasoning from images, which usually contains spatial relations between objects, we recommend the following techniques to enhance model capabilities:

- **Fine-tuning on Visuo-Spatial Tasks:** Developing and utilizing specialized fine-tuning datasets that explicitly target spatial reasoning skills, such as visual question answering (VQA) datasets focused on physics, engineering, requiring deep spatial understanding
- **Multi-modal Architectures with Enhanced Visual Encoders:** Exploring architectural improvements that allow for richer and more detailed encoding of visual information, potentially using more powerful vision backbones to capture fine-grained spatial details and relationships between visual elements.

To combat the issue with poor/faulty reasoning, which stems from hallucination, we propose the incorporation of domain-specific knowledge through a neuro-symbolic architecture where atomic facts about physical principles can be embodied in the system. This is an area we are pursuing with a hybrid architecture that combines fine-tuned LLM with automated reasoning systems to generate trustworthy, verifiable results with proof, accounting for spatial and commonsense reasoning, without compromising the system’s speed and robustness.

References

- [Achiam *et al.*, 2023] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- [Bai *et al.*, 2023] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023.
- [Banerjee *et al.*, 2022] Shreya Banerjee, Selmer Bringsjord, Michael Giancola, and Naveen Sundar Govindarajulu. Qualitative mechanical problem-solving by artificial agents: Further progress, under psychometric AI. In *The Int. FLAIRS Conf. Proc.*, volume 35, May 2022. doi: 10.32473/flairs.v35i.130630.
- [Banerjee, 2023] Shreya Banerjee. *Qualitative Mechanical Problem-Solving by Artificial Agents Using Hybrid AI*. PhD thesis, Rensselaer Polytechnic Institute, 2023.
- [Bennett, 2008] George K Bennett. Bennett Mechanical Comprehension Test. *Pearson*, 281:29, 2008.
- [Bringsjord and Schimanski, 2003] S. Bringsjord and B. Schimanski. What is artificial intelligence? Psychometric AI as an answer. In *Proc. of the 18th Int. Joint Conf. on Artif. Intell. (IJCAI-03)*, pages 887–893, San Francisco, CA, USA, 2003. Morgan Kaufmann.
- [Forbus, 1996] K.D. Forbus. Qualitative reasoning. In *CRC Handbook of Computer Science*, pages 715–733. Northwestern University, Evanston, IL, USA, 1996.
- [Hayes, 1978] Patrick J. Hayes. The naïve physics manifesto. In D. Mitchie, editor, *Expert Systems in the Microelectronics Age*, pages 242–270. Edinburgh University Press, Edinburgh, Scotland, 1978.
- [Hayes, 1985] Patrick J. Hayes. The second naïve physics manifesto. In J. R. Hobbs and B. Moore, editors, *Formal Theories of the Commonsense World*, pages 1–36. Ablex, Norwood, NJ, USA, 1985.
- [iPREP, 2022] iPREP. Bennett mechanical comprehension test ii (BMCT) — free practice & more, 2022. Accessed: Oct. 10, 2022.
- [Klenk *et al.*, 2011] Matthew Klenk, Kenneth D Forbus, Emmett Tomai, and Hyeonkyeong Kim. Using analogical model formulation with sketches to solve Bennett Mechanical Comprehension Test problems. *J. of Exp. & Theor. Artif. Intell.*, 23(3):299–327, Sep. 2011.
- [Team *et al.*, 2025] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.
- [Weld and de Kleer, 2013] Daniel S Weld and Johan de Kleer. *Readings in Qualitative Reasoning about Physical Systems*. Morgan Kaufmann, San Mateo, CA, USA, 2013.