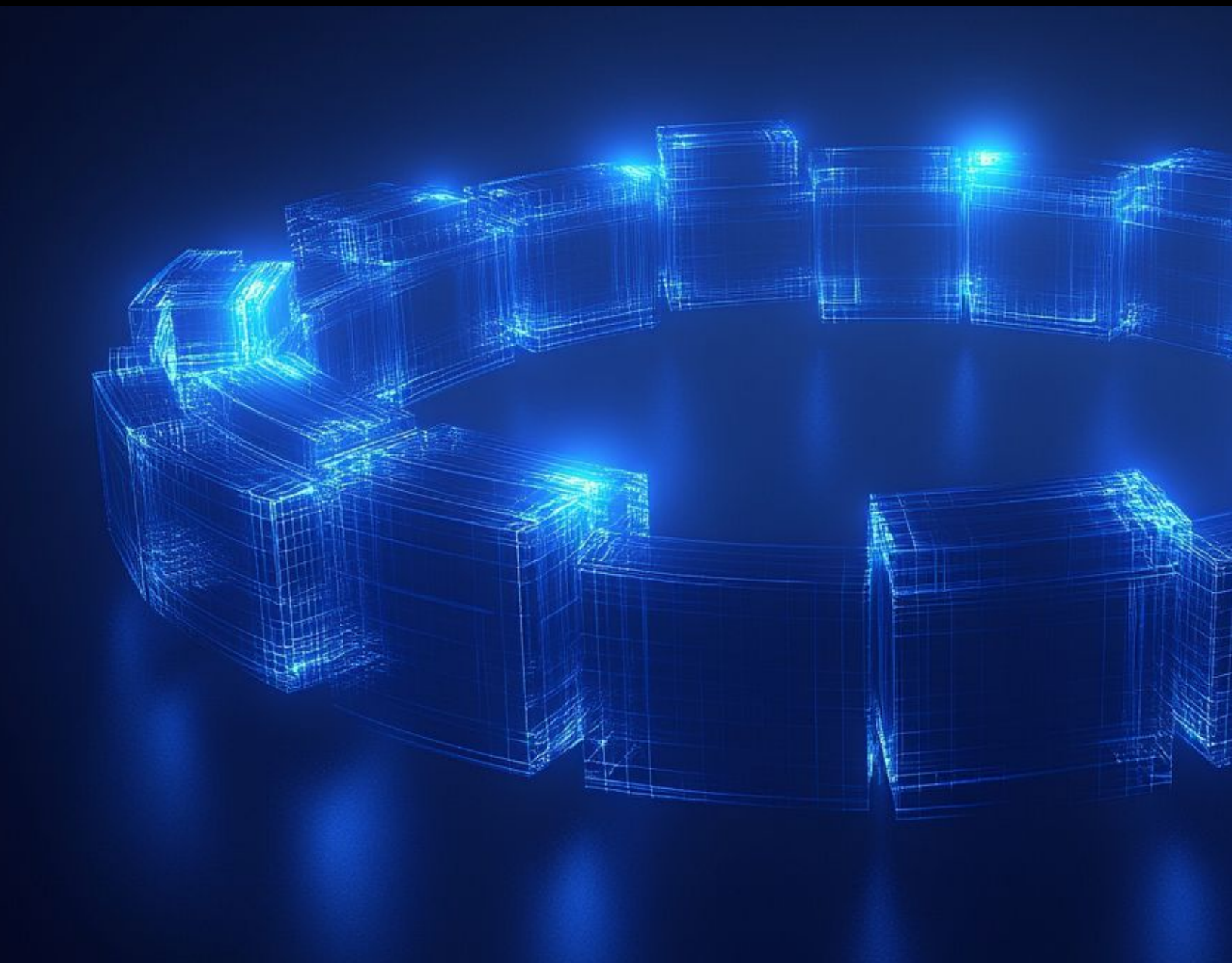
TURING

Turing VLM-Bench 1.0 for Real-World Tasks in Business and STEM

Advancing VLM evaluation with open-ended, multimodal tasks grounded in real utility



Introduction

Vision Language Models (VLMs) have shown remarkable progress in the past few years, guided by multiple vision-language benchmarks. While these benchmarks often capture broad or generalized tasks (like identifying everyday objects, answering everyday questions, or referencing common knowledge), they fall short in use cases where domain-specific data or technical jargon is critical. For instance, a *business professional* might need insights about finance, markets, or operational metrics, while a *chemist* might require detailed analysis of molecular structures or reaction pathways. Standard benchmarks rarely test these real-world scenarios, creating a gap between benchmark scores and actual usefulness in professional workflows.

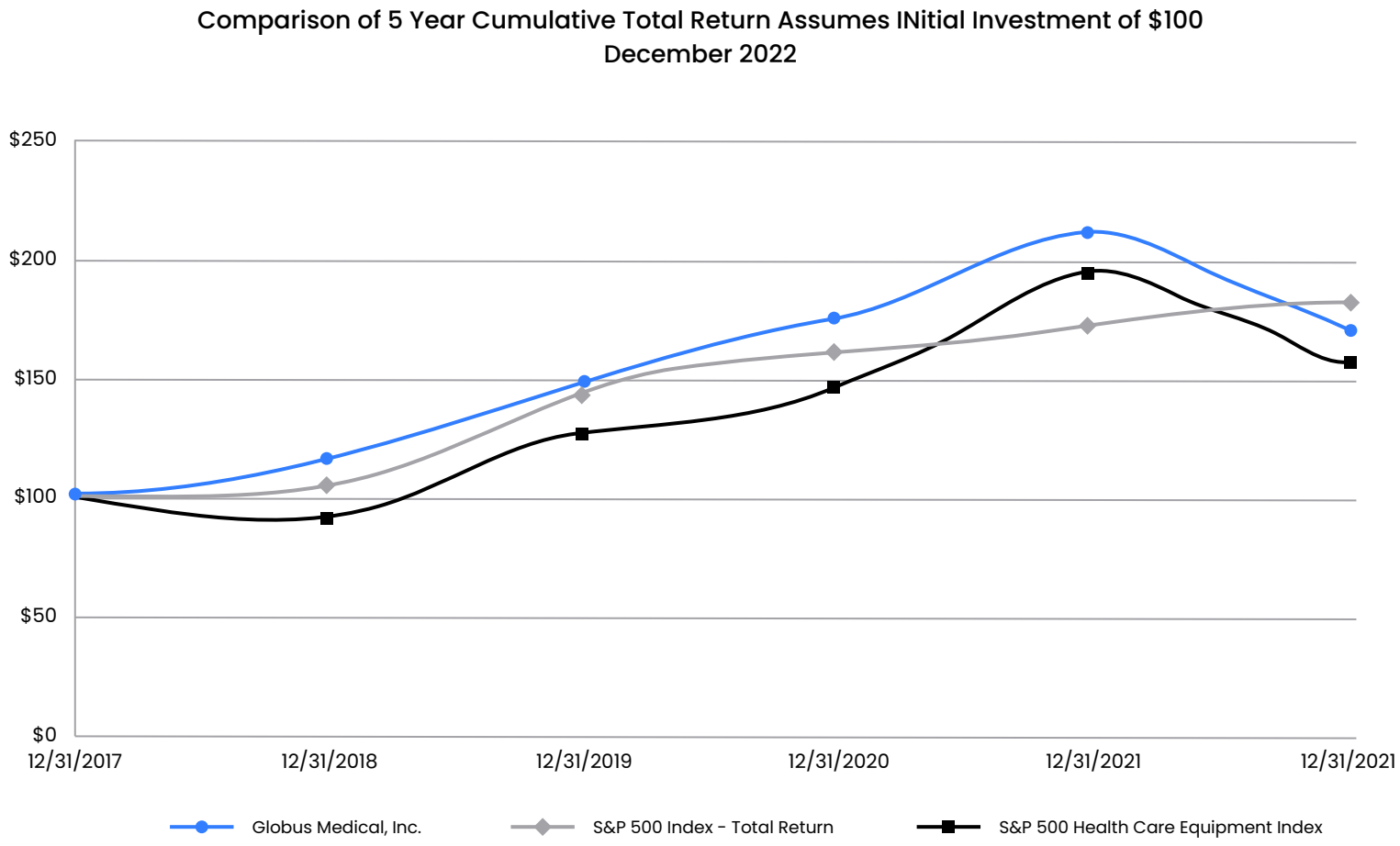
Our new benchmark is uniquely driven by the need to assess VLM performance on tasks that matter in real-world workflows. These tasks are designed to test whether a model can apply conceptual knowledge to realistic, workflow-driven queries, rather than artificial prompts created just to quiz the model. Specifically, each answer is intended to support a downstream decision in the user's workflow. To accomplish this, we focus on **realistic vision-language scenarios** drawn from **business and STEM domains**, featuring relevant content that reflects the practical challenges professionals face. The benchmark includes text+image inputs (e.g., documents with tables, charts, or diagrams) and requires the model to generate open-ended responses rather than selecting from multiple-choice options. This setup forces models to reason and generate free-form answers in a format aligned with real-world decision-making workflows. The accuracy of each answer is assessed by a large language model **(LLM)-based judge**, allowing for nuanced evaluations. This approach benefits both AI researchers, who can advance their work on complex, high-stakes tasks, and industry professionals, who need to determine how well these models handle actual domain-specific problems. The following sections detail the benchmark's construction, methodology, findings, and future directions.

Feature Comparison: Turing-VLM Bench 1.0 vs. Public Vision Benchmarks

Benchmark	Primary Focus	Data Collection Method	Answer Format	Evaluation Approach	Unique Differentiators
Turing VLM Bench 1.0	Real-world professional tasks (e.g., structural engineering)	Curated by domain experts	Open-ended factual phrases	LLM-as-Judge for answer verification	Emphasis on practical, professional applications; expert-authored Q&A pairs
MMMU	Multidisciplinary academic knowledge across various fields	Sourced from college exams and textbooks	Multiple-choice questions	Standard accuracy metrics	Broad coverage of academic disciplines; diverse image formats
VisIT-Bench	Instruction-following in real-world inspired scenarios	Human-authored instructions and captions	Instruction-conditioned captions	Automated comparison using GPT-4 and human evaluations	Focus on instruction-following capabilities; diverse task range
@Bench	Assistive technology tasks for visually impaired users	Guided by user studies with people with visual impairments	Task-specific outputs (e.g., captions, OCR results)	Task-specific evaluation metrics	Tailored for assistive technologies; multi-task evaluation
ScienceQA	Multimodal science question answering	Collected from elementary and high school science curricula	Multiple-choice with explanations	Accuracy and chain-of-thought reasoning	Includes lectures and explanations; emphasizes reasoning processes
MathVista	Mathematical reasoning in visual contexts	Aggregated from 28 datasets + 3 new ones (IQTest, FunctionQA, PaperQA)	Multiple-choice and open-ended	Accuracy and reasoning depth	Evaluates algebraic, geometric, and scientific reasoning with visual data
RealWorldQA	Spatial understanding in real-world images	Images captured from vehicles and other real world scenarios	Multiple-choice and open-ended	Manual verification and accuracy metrics	Focus on real-world spatial reasoning; high-resolution images

Benchmark Details

Our benchmark consists of a diverse set of **vision-language prompts** that pair textual context with related images. These images include business graphics (like financial charts, sales graphs, or organizational diagrams) and STEM visuals (like scientific charts, engineering schematics, or data tables). Each prompt is carefully designed to reflect a realistic scenario—for instance, analyzing a quarterly revenue chart, interpreting experimental results from a lab report, or extracting insights from a complex diagram. The text portion of a prompt often provides background information or specific questions, while the image provides data that must be interpreted. This **multimodal setup** requires models to **read and cross-reference both textual and visual content**, a task requiring perception, domain knowledge, and reasoning. This interleaving of text and images mirrors real tasks (e.g., a financial analyst reading a report with charts, or a scientist reviewing a study with graphs), making the benchmark problems challenging and relevant.



[Image source]

Example prompt: A hedge fund matched Globus Medical's (GMED) 81% return from 2017-2022 but charged a 2% annual management fee (compounded) and a 20% performance fee on profits. Getting other details from the image, help me determine by how much the net return after fees in the hedge fund would outperform a no-fee S&P 500 investment?

In this sample task, the model is tasked with responding to a financial analyst's question on how to estimate the net return difference between the hedge fund and S&P 500 Index Fund. To answer correctly, the model must **interpret the diagram, track quarterly trends across curves, and apply financial reasoning using domain knowledge.** This single prompt thus tests multiple skills: visual comprehension of the diagram, integration with the

Benchmark Details (cont.)

textual question, and financial reasoning to solve the problem. In another scenario, a prompt might include an image of a data table from an annual report and a question about percentage changes; the model would need to **extract the right figures from the table image** and calculate the change. By covering such scenarios, the benchmark tasks reflect the complexity of real-world problems—from understanding trends in business metrics to comprehending scientific data—rather than the simplified queries found in many academic datasets.

Key VLM Capabilities Tested

The overarching goal of the benchmark is to provide a multidimensional metric for strengths and weaknesses of VLMs. We have designed the benchmark to pressure-test VLMs across the following well-known areas of difficulty:

- **Advanced perception**: Ability to accurately identify and interpret relevant visual elements or details within complex or information-rich images. Examples: identifying a specific financial metric from a cluttered earnings report or accounting statement, or recognizing key components within a detailed chemical diagram or biological illustration.

- **Spatial reasoning**: Ability to interpret, understand, and reason about spatial relationships, arrangements, or configurations in visual data. Examples: determining how different parts in an engineering blueprint fit together, or interpreting circuit diagrams to understand the configuration (series vs. parallel) of electrical components.

- **Numerical reasoning**: Ability to extract numerical information from visual data, perform quantitative comparisons or calculations, and reason about numerical trends or relationships. Examples: calculating revenue growth percentages from bar charts of quarterly sales, inferring numerical relationships from graphs showing chemical concentrations or biological population growth.

- **Logical inference**: Ability to draw conclusions, deduce implications, or predict outcomes based on multimodal data involving logical reasoning and cause-effect relationships. Examples: inferring the economic outcome given graphical data on unemployment rates and consumer spending, predicting the next step in a chemical reaction sequence depicted in a visual diagram, or reasoning about cause-and-effect in physics or chemistry experiments.

- **Temporal reasoning**: Understanding sequences, timelines, or changes over time. Example: interpreting economic trends from a time-series plot, or predicting the next state in a biological process.

- **Contextual commonsense reasoning**: Applying common sense or general knowledge in contextually rich scenarios. Examples: interpreting a business scenario (e.g., identifying unusual expenses), or understanding implicit information in engineering diagrams (e.g., inferring that a structure might fail under certain conditions).

- **Abstract or analogical reasoning**: Drawing parallels between visually different but conceptually similar scenarios. Examples: comparing patterns of population growth graphs in biology to similar trend graphs in economics, or interpreting engineering diagrams by analogy to well-known real-world systems.

- **Counterfactual reasoning:** Reasoning about hypothetical scenarios or what-if questions. Examples: predicting outcomes if a specific financial parameter changed ("What if interest rates doubled?"), or hypothetical scenarios in physics ("What if friction were eliminated from this system?").

- **Iterative or multi-step reasoning**: Solving multi-step problems or reasoning tasks that require iteratively interpreting visual data. Example: interpreting multiple layers of diagrams or charts sequentially to answer a final inference question.

Dataset Description

Turing VLM-Bench 1.0 includes 705 tasks of varying complexity across subjects in business and STEM. We started with a larger set of questions, which annotators tested against multiple state-of-the-art (SOTA) VLMs. Questions were selected only if they were consistently answered incorrectly by at least one of the evaluated SOTA models. We refer to this full set of filtered questions as the "**ALL**" set in the remainder of this report.

We also created a **HARD** subset by removing tasks where the average accuracy across any SOTA model exceeded 50%, as measured by our evaluation script.

Below we report detailed statistics for both our **ALL** and **HARD** sets.

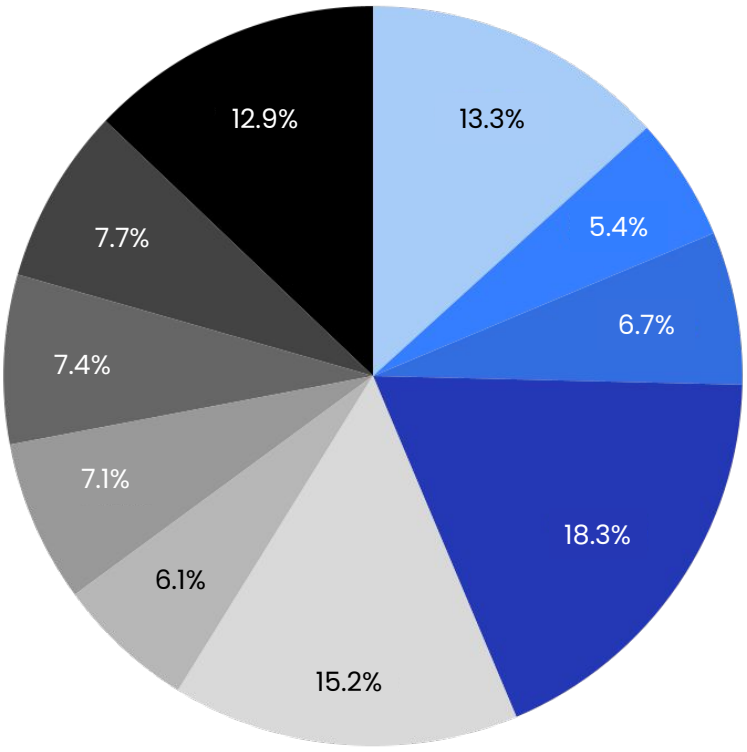
Data Statistics

To support a consistent annotation process, we developed a two-level taxonomy (L1 and L2) covering subjects and task types across business and STEM domains. The task distribution across taxonomy categories is shown below.

Task Distribution

ALL set

L2 Task Category Distribution (Benchmark Set)



L1 Category - STEM (n=397)
Math (n=91)
Electrical and Electronics Engineering (n=54)
Civil and Structural Engineering (n=52)
Computer Engineering and Software Engineering (n=50)
Mechanical and Aerospace Engineering (n=43)
STEM - Others (n=107): Physics (21), Artificial Intelligence (20), Chemical and Process Engineering (17), Other Engineering (12), Chemistry (11), Earth Sciences (e.g. Geology, Hydrology, Meteorology, Environmental Science) (8), Information Technology (7), Communication and Networking (6), Emerging and Specialized Tech (2), Biology (2), Other Specializations (1)

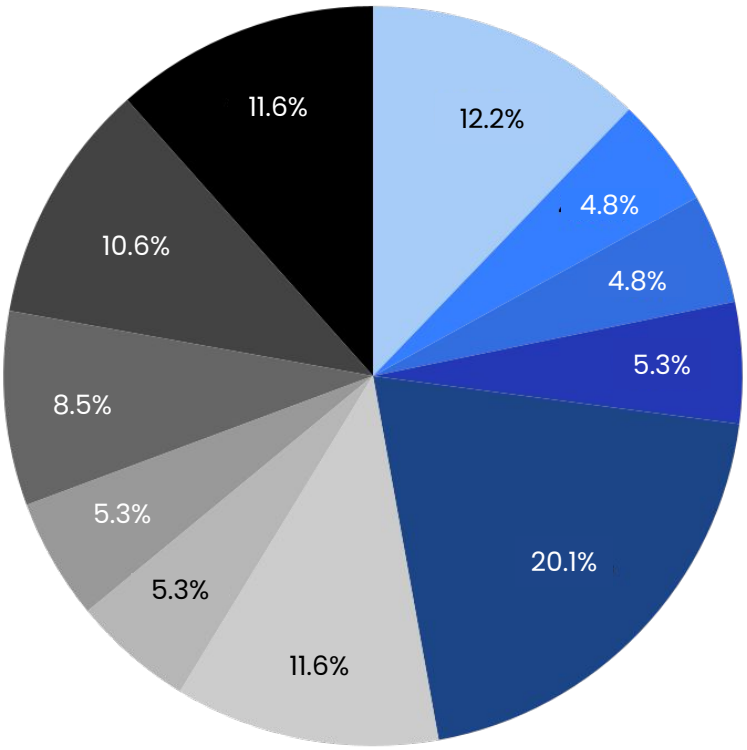
L1 Category - Business (n=308)
Finance (n=129)
Marketing and Sales (n=47)
Business Analytics and Information Systems (n=38)
Business - Others (n=94): Management & Leadership (26), Operations and Supply Chain (20), Accounting (18), Economics and Global Business (14), Human Resources (HR) (12), Entrepreneurship and Innovation (4)

Data Statistics (cont.)

Task Distribution

HARD subset

L2 Task Category Distribution (Hard Benchmark Set)



L1 Category - STEM (n=100)

- Math (n=22)
- Electrical and Electronics Engineering (n=20)
- Civil and Structural Engineering (n=16)
- Computer Engineering and Software Engineering (n=10)
- Mechanical and Aerospace Engineering (n=10)
- STEM - Others (n=22): Chemical and Process Engineering (5), Physics (4), Artificial Intelligence (4), Other Engineering (3), Earth Sciences (e.g. Geology, Hydrology, Meteorology, Environmental Science) (2), Chemistry (1), Biology (1), Communication and Networking (1), Other Specializations (1)

L1 Category - Business (n=89)

- Finance (n=38)
- Management and Leadership (n=10)
- Business Analytics and Information Systems (n=9)
- Marketing and Sales (n=9)
- Business - Others (n=23): Operations and Supply Chain (7), Economics and Global Business (6), Accounting (6), Human Resources (HR) (3), Entrepreneurship and Innovation (1)

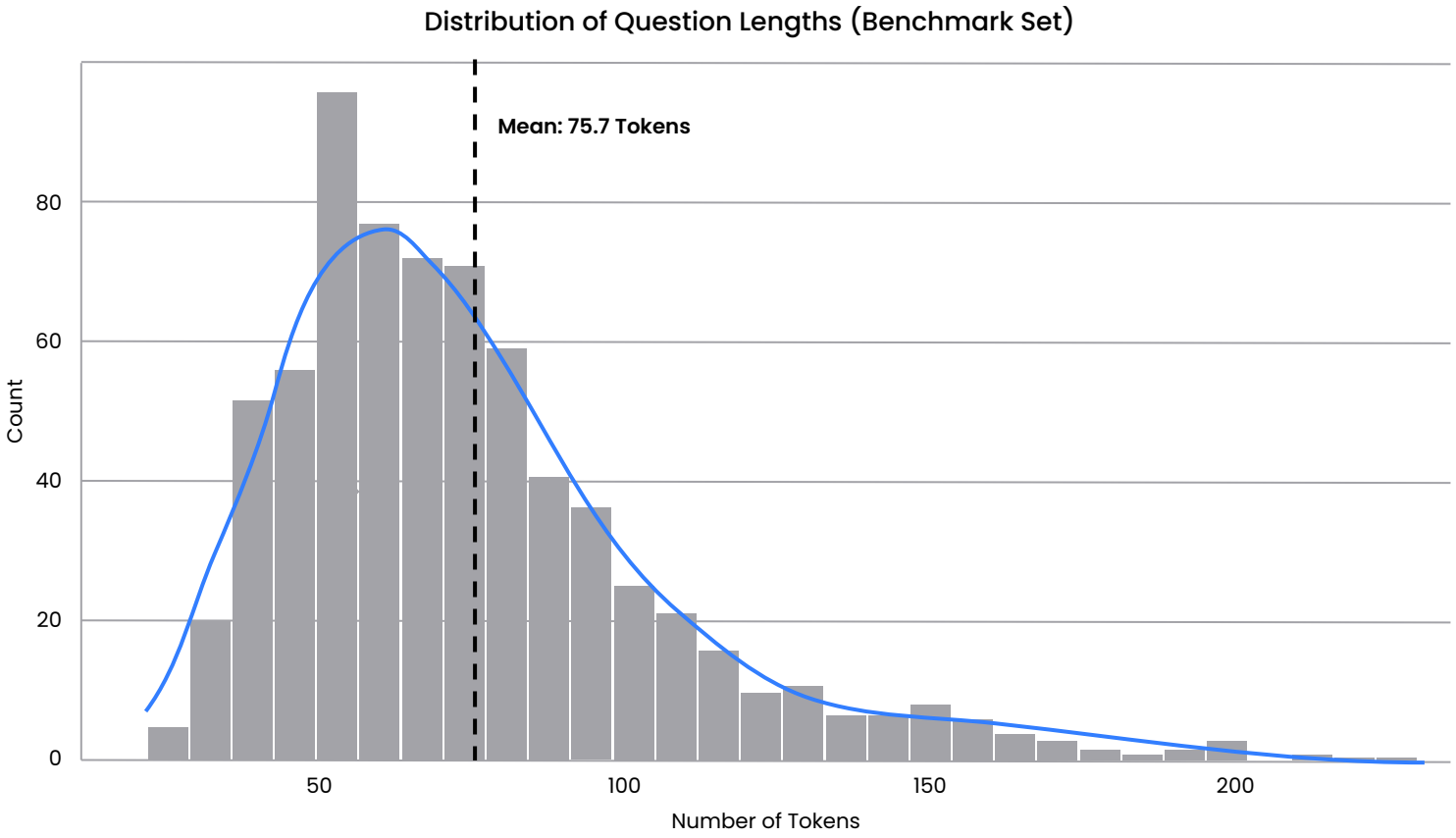
Question Lengths

We analyzed the distribution of question lengths, as both length and complexity can significantly affect model performance:

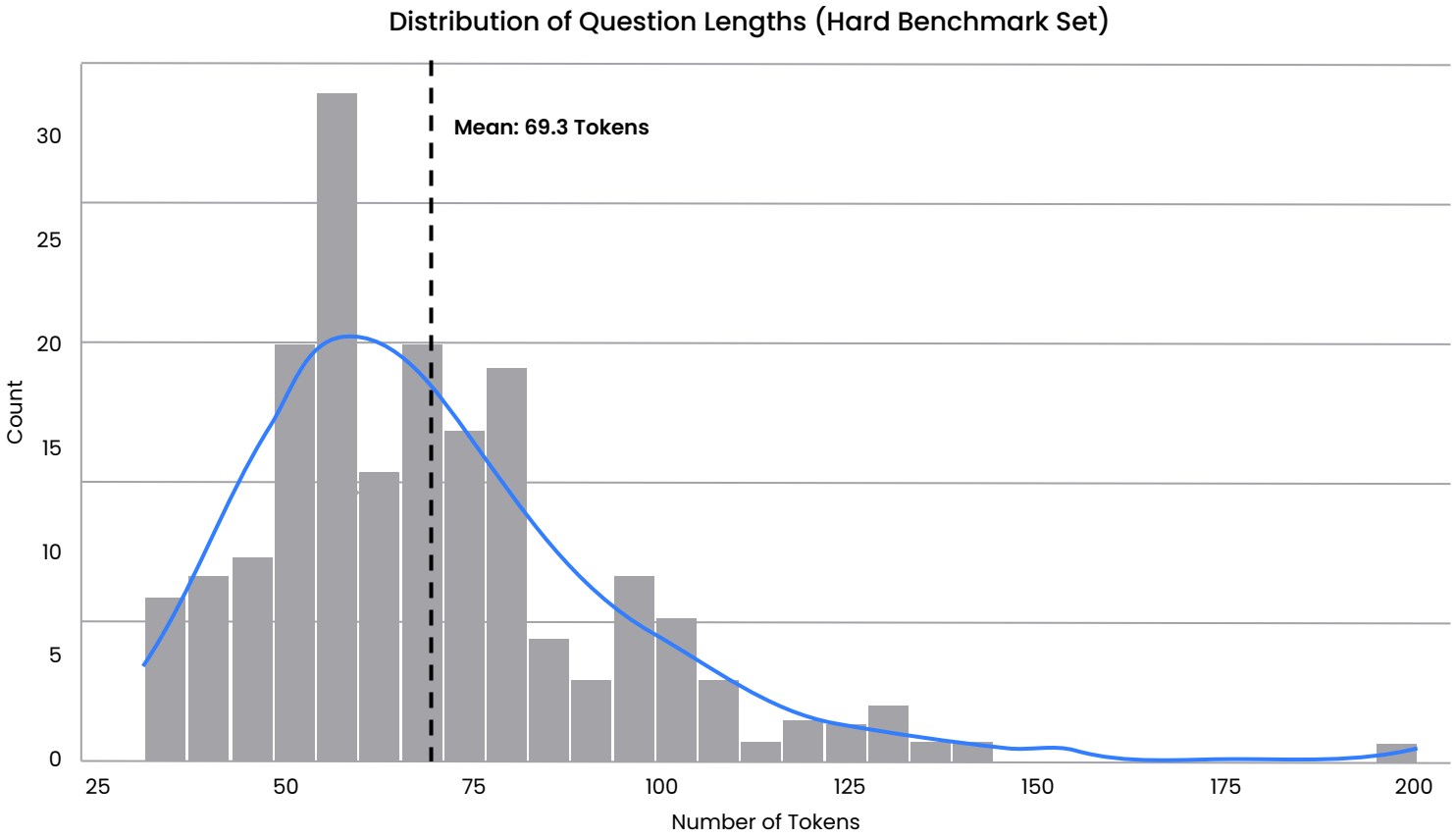
- **Complexity and reasoning:** Longer, more detailed questions often require complex reasoning and multi-step inference. Vision models need to parse the text, identify key information, and integrate it with visual data. Longer questions may increase the cognitive load on the model, potentially leading to errors or less accurate responses.
- **Context understanding:** While longer questions offer more context, they also require the model to maintain and utilize this context effectively. If a model struggles with long-range dependencies or maintaining context, it may fail to answer correctly.
- **Data extraction:** Questions that require extracting specific details from images or tables become more challenging when embedded in a longer question. The model needs to filter out irrelevant information and focus on the target data.

Data Statistics (cont.)

ALL set



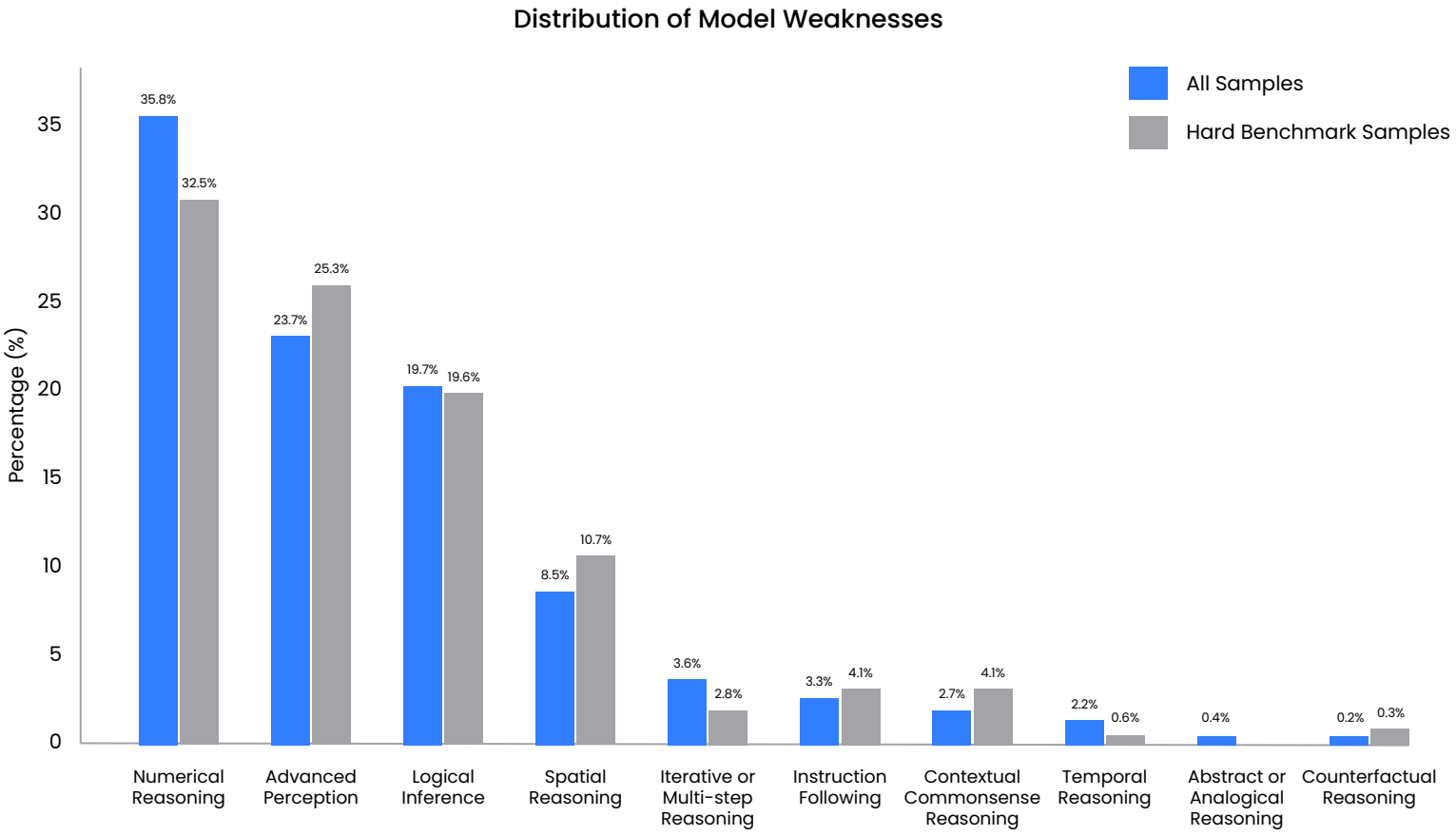
HARD subset



Data Statistics (cont.)

VLM Capabilities Tested

The following chart shows the distribution of the various VLM capabilities tested across the samples in the benchmark. Note that any given task in the benchmark can test for multiple capabilities of the VLM.



Annotation Process

This section outlines the rigorous, multi-stage annotation process behind the benchmark's creation.

Skilled and Experienced Annotators

The creation of our high-quality benchmark began with assembling a team of qualified professionals with formal training and experience in relevant disciplines. Our annotators possess diverse academic and professional backgrounds with practical industry experience that complements their theoretical knowledge. Team members hold advanced degrees and/or have relevant work experience in fields including business management, sciences, engineering, and data science, ensuring that the prompts they create or review demonstrate real-world relevance and technical accuracy. All annotators completed a comprehensive selection process involving structured interviews and proctored assignments. Following recruitment, candidates participated in a standardized training program; only those who exceeded established performance thresholds qualified to contribute to this project.

Task Distribution Based on Expertise

We recognize that different domains demand different skill sets. To effectively balance complexity and clarity across varied subject areas, we assigned tasks to annotators based on their proven expertise:

- **Technical and STEM domains:** Annotators with engineering, physics, or computer science backgrounds were assigned tasks that required domain knowledge—such as reading circuit diagrams, describing chemical structures, or analyzing code snippets.

- **Business and finance:** Annotators with professional experience in finance, accounting, or marketing were assigned tasks that involved interpreting stock charts, business reports, or market trends.

While practical utility was the benchmark's primary focus, we also ensured coverage across diverse foundational skills required to solve each task. The question prompts were designed to test multiple types of questions, including but not limited to direct questions, conditional questions, reasoning by analogy questions, and estimation-based questions.

By diversifying question formats, the benchmark can reveal how the model handles instructions and answers in different structures, and ensure that weaknesses are not masked by a particular question style. *Any prompt-image pair that produced invalid responses from one or more leading SOTA VLMs was included in the final benchmark.*

Annotation Process (cont.)

Quality Control

Maintaining the integrity of any benchmark data requires rigorous oversight. Our approach involved a multi-layered review strategy to ensure that each prompt-image pair met high standards of clarity, correctness, and relevance:

1. **Three-step professional review**

    a. **Initial screening:** Prompt-image pairs were examined by a team of generalist reviewers equipped with clear guidelines. They evaluated aspects such as prompt clarity, alignment of prompt to image, and basic factual correctness in both the request and the purported "ideal response."

    b. **Domain expertise check:** Samples that passed the first screening were then reviewed by subject matter experts in relevant business areas or STEM fields. Their specialized knowledge allowed them to validate the technical or domain-specific content of the prompts and responses. For instance, prompts about strength of materials would be vetted by a reviewer with a background in mechanical engineering, ensuring the correctness and plausibility of the related images and answers.

    c. **Final approval:** In the last stage of professional review, experienced editors re-examined the samples to confirm that earlier feedback had been effectively addressed. They were responsible for spotting inconsistencies, ensuring compliance with established style guides, and verifying that the prompts were neither ambiguous nor overly simplistic.

2. **Researcher spot checks**

    After passing through the three-layered review, selected samples were subjected to an additional "spot check" by the research team responsible for creating and curating the benchmark. These researchers performed random audits to confirm the quality and reliability of the dataset. Their insights often led to improvements such as refining specialized prompts, updating references in the ideal response, or flagging subtle image discrepancies that might have been overlooked.

3. **Domain-specific validation**

    Given that our prompts frequently span specialized fields—from business to engineering and data science—only reviewers and researchers with proven domain knowledge were permitted to validate the corresponding items. This ensured that both the prompt and the expected answer held up to real-world scrutiny in those fields. As a result, the dataset remains credible and practical for evaluating vision-language models in those specific domains.

Annotation Process (cont.)

Common Reasons for Rejection

Despite meticulous planning, certain prompts and images did not meet our quality benchmarks. The primary reasons for rejection included:

1. **Contrived or unclear prompts:** If the prompt was deemed too artificial, trivial, or overly convoluted without adding real analytical value, it failed to provide meaningful evaluation data.

2. **Poor-quality images:** Blurry, low-resolution, or otherwise unusable images risked skewing model performance and were rejected to maintain a consistent level of visual clarity.

3. **Lack of uniqueness:** Duplicates or near-duplicates offered limited incremental benefit to the dataset, so they were removed to ensure diverse coverage of contexts and concepts.

4. **Inaccurate ideal response:** Any dissonance between the prompt and its ideal solution—especially if factual or domain-specific errors were present—led to immediate rejection or a request for revision.

By enforcing these robust protocols, we ensured that the benchmark is both comprehensive and reliable, so that vision-language models trained or tested against this dataset face realistic, high-quality evaluation scenarios. This detailed curation process underscores our commitment to creating a rigorous standard for performance assessment in the rapidly evolving field of multimodal AI.

Evaluation

Evaluating VLMs requires a structured, end-to-end approach—from framework selection to robust scoring protocols.

Framework

We examined multiple evaluation frameworks and decided to adapt vlmevalkit for its broad support of multimodal data formats and flexible model integration—via API or local deployment. This paper provides the detailed design of the vlmevalkit framework.

Method

Evaluating open-ended responses requires a different approach than standard multiple-choice quizzes. We opted for an LLM-as-a-judge strategy, where an LLM acts as an automated evaluator of answers. This method has gained popularity as a practical alternative to costly human evaluation for free-form text (A Survey on LLM-as-a-Judge). It offers flexibility: instead of exact string matching or limited multiple-choice keys, the AI judge can consider whether the answer is essentially correct even if phrased in a novel way. However, LLM-as-a-judge introduces potential limitations around bias and consistency. To address this concern, we analyzed the variance in our scoring approach due to the stochastic nature of evaluation using LLM-as-a-judge and ensured that it is minimal. We also cross-verified LLM-as-a-judge results with human experts on a subset of 300 examples in our benchmark and found disagreement in only 1% of examples.

Our evaluation pipeline involves multiple steps to ensure a robust and fair scoring:

1. **Multiple independent runs:** For each prompt, we generate **five independent model responses**. LLMs can produce different outputs on different runs due to their probabilistic nature, so this step captures the variability in the model's performance on a given question.

2. **LLM judging each response:** An AI judge (a strong LLM configured for evaluation) reviews each response and **assigns an accuracy score between 0 and 1**. A score of 1.0 means the answer is completely correct, 0.5 might indicate a partially correct answer, and 0 means the response is entirely incorrect. The judge's prompt is crafted to consider the factual correctness and completeness of the answer against the question's requirements. In our implementation, the judge has access to the prompt and the reference solution or criteria for correctness to base its evaluation on.

3. **Averaging scores:** We calculate the **final accuracy for the prompt as the average of the five runs' scores**. This averaged score smooths out randomness and gives a more stable estimate of the model's true performance on that prompt. For example, if a model sometimes gets an answer right and other times wrong, its average might be 0.5, reflecting inconsistent knowledge or understanding.

This judge-based evaluation has several advantages. **Unlike multiple-choice tests, there's no opportunity for the model to guess from given options or be biased by how the options are written**. Multiple-choice questions, while convenient for auto-grading, can introduce their own biases—e.g., LLMs have shown sensitivity to the order of choices and can even be gamed by simple heuristics (*Gaming TruthfulQA: Simple Heuristics Exposed Dataset Weaknesses*). By requiring the model to produce an answer from scratch, we force it to fully engage with the question. The LLM judge can then award partial credit when appropriate, something a strict correct/incorrect metric would miss. This means our evaluation **captures the model's reasoning quality better**—a model that arrives at a nearly correct conclusion may get 0.8, whereas one that is totally off-base gets 0, with shades in between.

Results

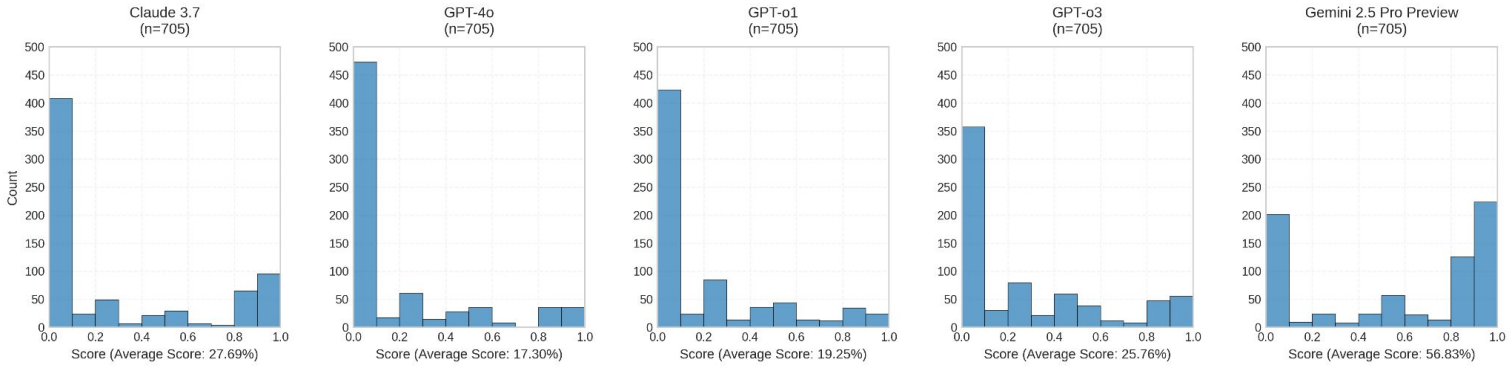
Below, we report results for four leading VLMs across both the ALL and HARD subsets of the benchmark.

1. Overall accuracy (average score) along with its 95% confidence interval
2. Distribution of model scores over the subset

ALL
Benchmark Dataset

Model	Accuracy (%)	95% Confidence Interval (%)
Gemini 2.5 Pro Preview	56.83	[53.8%, 59.87%]
Claude 3.7	27.69	[24.85%, 30.53%]
GPT-o3	25.76	[23.32%, 28.2%]
GPT-o1	19.25	[17.11%, 21.38%]
GPT-4o	17.3	[15.11%, 19.5%]

Distribution of Average Score

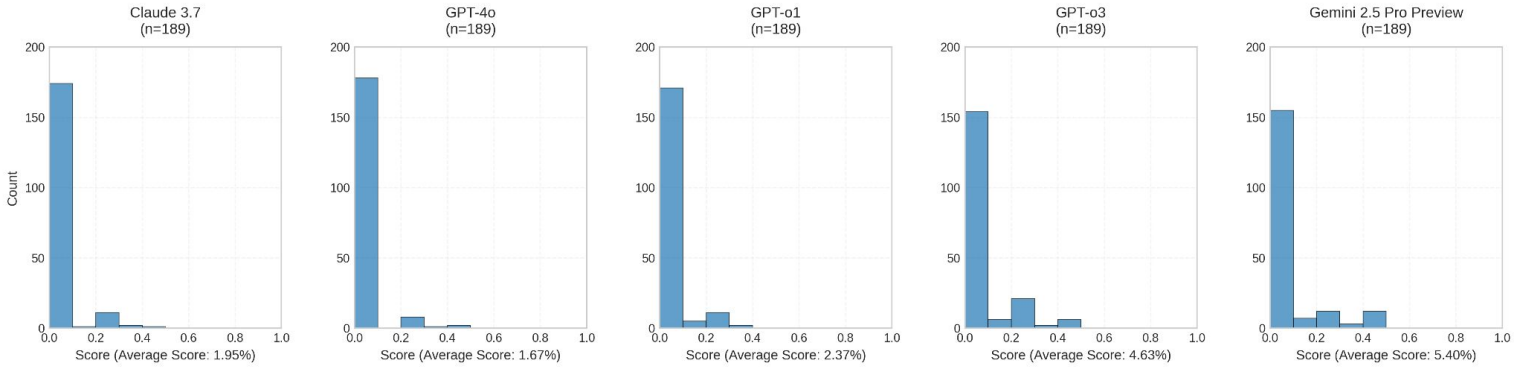


Results

HARD
Hard Benchmark Dataset

Model	Accuracy (%)	95% Confidence Interval (%)
Gemini 2.5 Pro Preview	5.4	[3.67%, 7.12%]
GPT-o3	4.63	[3.18%, 6.09%]
GPT-o1	2.37	[1.46%, 3.28%]
Claude 3.7	1.95	[0.98%, 2.92%]
GPT-4o	1.67	[0.75%, 2.59%]

Distribution of Average Score



Error Analysis

To gain deeper insight into a model's weaknesses and strengths, especially in core vision-related competencies, we computed an **average capability score** for each foundational VLM capability across the entire benchmark. Here's how we approached it:

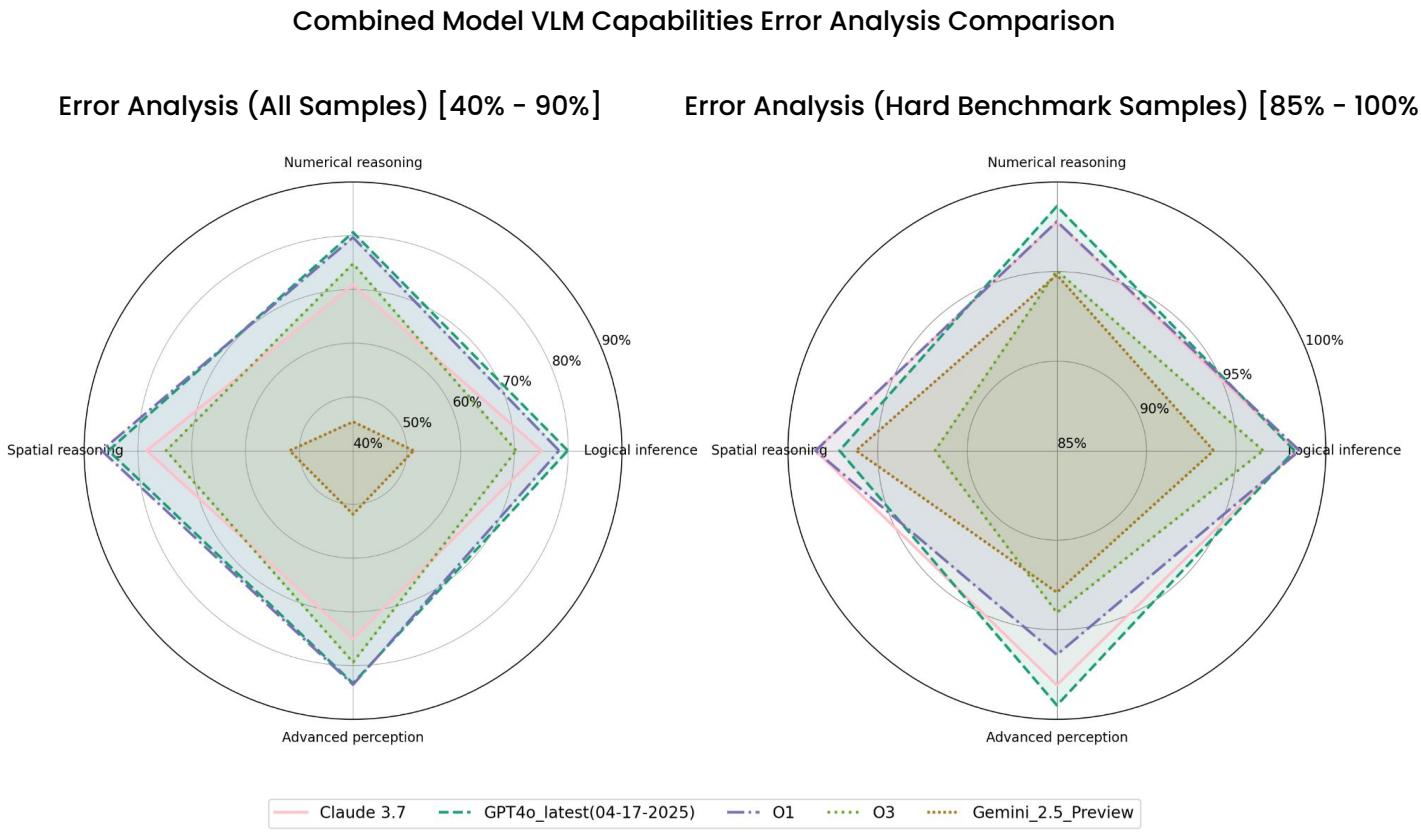
1. **VLM capability tagging and partial scores**
   Each question in our dataset was labeled with one or more **foundational VLM capabilities,** such as "Advanced perception," "Spatial reasoning," or "Logical inference." After the model produced its response, we evaluated the correctness of the answer and, where appropriate, assigned **partial credit** for partially correct or incomplete responses.

2. **Aggregated performance for each VLM capability**
   We then aggregated these partial (or full) scores across all items tagged with the same VLM capability. By averaging the model's performance for each capability, we derived a single "capability score" that reflects how well the model handles questions involving that particular foundational capability.

Results (cont.)

3. **Error chart vs. benchmark frequency**
   Next, we **plotted an error chart** to visualize how often a given capability appeared in the benchmark (its frequency) alongside the model's average score for that capability. This approach reveals not only where the model may struggle but also how critical that capability is relative to the rest of the test set. If a capability occurs frequently and the model's corresponding score is low, it underscores a potentially serious weakness that could impact overall performance. To reduce noise, we focused analysis on the four most represented capabilities in the dataset.

   The chart below illustrates each model's weakness profile across various capabilities.



Key takeaways from this analysis include:

- **Gemini-2.5-Preview** consistently outperforms the other models in the overall benchmark set.
- Across the full benchmark, the largest performance gaps appear in **spatial reasoning** and **advanced perception**, suggesting that accuracy improvements in these domains are prerequisites for further progress in vision models.
- Model errors remain generally uniform on the hard subset of the benchmark, underscoring the challenge of addressing these advanced tasks, which require more robust reasoning and domain-specific capabilities to achieve significant performance gains.

Future Work

Looking ahead, there are several directions to enhance the benchmark and its evaluation methodology:

- **Expand and diversify the benchmark:** We plan to broaden the range of domains and problem types. This could include more fine-grained **STEM subjects** (e.g. biology lab reports, physics diagrams) and additional **business scenarios** (like finance, marketing, or operations data). We also intend to incorporate more image types, such as scanned forms, blueprints, and even combinations of multiple images, to better test model perception and compositional reasoning. Expanding the dataset in this way ensures models are evaluated across a broad spectrum of real-world challenges—reducing reliance on pattern memorization and emphasizing applied reasoning.

- **Refine the evaluation methodology:** While the LLM-as-a-judge approach works well, we are exploring ways to make it even more robust and informative. One idea is to have the judge provide a structured explanation or **rubric-based scoring** for each response, breaking down the score into categories (e.g., correctness of reasoning, accuracy of data extraction, etc.). This would give more detailed feedback on *why* a model's answer was right or wrong. We're also addressing the LLM judge's reliability—focusing on consistency and reducing bias to ensure scoring remains credible at scale. (A Survey on LLM-as-a-Judge).

- **Industry collaboration and continuous improvement:** We invite collaboration from industry partners facing real-world multimodal challenges, whose input can help make prompts more realistic and keep the benchmark aligned with evolving use cases. For example, a finance company might contribute anonymized examples of analyzing annual reports, or a healthcare organization might suggest prompts involving medical charts. By working together with domain experts, we can iteratively improve the benchmark's coverage and difficulty. We also plan to periodically evaluate new models (and new versions of models) on the benchmark and its future enhanced versions. This **continuous evaluation** will track progress over time and highlight where breakthroughs are happening or where models still fall short. Although the benchmark is private, we're considering mechanisms to allow external researchers to test their models on it (e.g., via a limited-access leaderboard or partnership program) so that the broader AI community can benefit from its insights.

Conclusion

We introduced the first version of a novel vision-language benchmark that stresses realistic prompts in Business and STEM, aiming to push AI systems closer to real-world competency. Through a combination of text and image inputs, these tasks demand the kind of comprehensive understanding and reasoning that traditional benchmarks often fail to capture. By replacing multiple-choice questions with open-ended prompts evaluated by an LLM judge, we obtain a richer picture of model performance—not just accuracy, but the ability to consistently generate correct, well-reasoned outputs.

Initial results show that while cutting-edge VLMs have made impressive strides, there remains a significant gap between current capabilities and the expert-level performance required for real-world business analytics or scientific reasoning.

The use of an LLM judge with a nuanced scoring system proved effective in highlighting partial knowledge and reasoning errors, offering valuable guidance for model improvement. We believe that realistic evaluation is a catalyst for innovation: by identifying exactly where models struggle, researchers and engineers can target those weaknesses with better training data, model architectures, or prompting techniques. For researchers and industry professionals alike, we hope this benchmark becomes both a practical tool and a foundation for the next wave of multimodal AI development.

We'll continue refining the benchmark and evaluation pipeline, sharing future results through new releases and potentially a collaborative platform. Progress in AI depends on benchmarks that reflect the complexity of the world AI is meant to serve, and we're excited to drive that progress together.

About Turing

Turing is one of the world's fastest-growing AI companies accelerating the advancement and deployment of powerful AI systems.

It helps customers in two ways: Working with the world's leading AI labs to advance frontier model capabilities in thinking, reasoning, coding, agentic behavior, multimodality, multilinguality, STEM and frontier knowledge; and leveraging that work to build real-world AI systems that solve mission-critical priorities for companies.

Turing—based in San Francisco, California—was named #1 on The Information's annual list of "Top 50 Most Promising B2B Companies," and has been profiled by Fast Company, TechCrunch, Reuters, Semafor, VentureBeat, Entrepreneur, CNBC, Forbes, and many others. Turing's leadership team includes AI technologists from Meta, Google, Microsoft, Apple, Amazon, X, Stanford, Caltech, and MIT.

Want to benchmark your VLM on real-world business and STEM tasks—and see how it stacks up against SOTA models?

Talk to a VLM expert to explore real-world evaluations or request sample data.