Design Documentation Decoded:

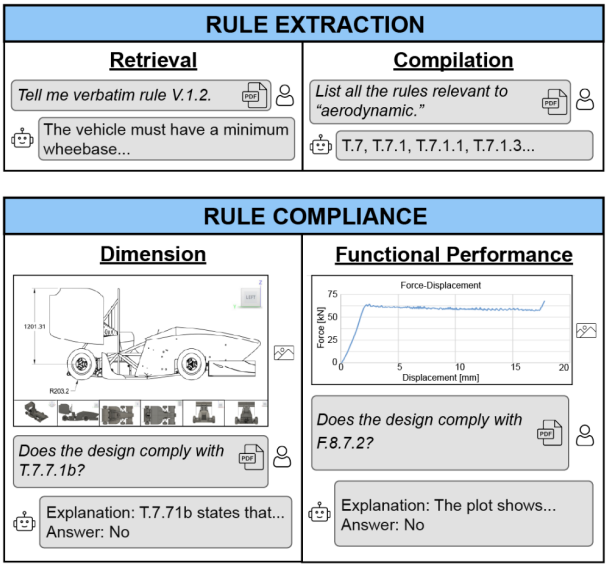
Improving AI's Understanding of Engineering Documents

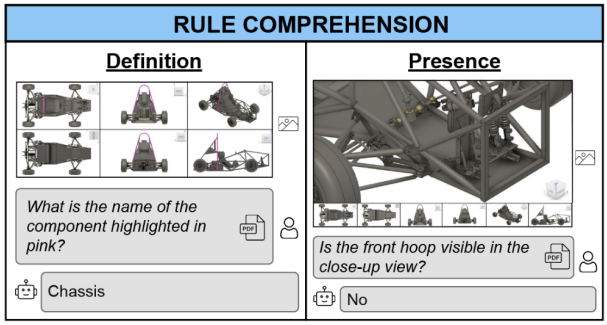
[DesignQA Benchmark]

Problem Statement

How well can LLMs understand design engineering documents?
Will model hallucinations exacerbate design biases and lead to design catastrophes?
Can you develop a method to beat the best-performing model on an engineering design benchmark, DesignQA?

A trusted AI assistant that can deeply understand technical engineering documents would be invaluable for engineers, especially for tasks that require frequent reference to standards or compliance checks. Unfortunately, research shows that vision-language models (VLMs) currently struggle with synthesizing technical information from documents with engineering images. To evaluate these capabilities, the DesignQA benchmark—composed of 1149 question-answer pairs based on real data from the MIT Motorsports team—tests VLMs on their ability to interpret and apply the Formula SAE rulebook. GPT-4o, Gemini 1.0, Claude Opus, and other models have been evaluated on DesignQA. However, no existing model achieves a perfect score on any DesignQA subset. The results for many of the benchmark subsets indicate that there is substantial room for VLM improvement when it comes to understanding and cross-analyzing engineering documentation and images.





Section	Subset	Metric	# QAs
Extraction	Retrieval	F1 BoW	1192
	Compilation	F1 Rules	30
Comprehension	Definition	F1 BoC	31
	Presence	ACC	62
Compliance	Dimension	ACC/Bleu/Rouge	120
	Functional Performance	ACC/Bleu/Rouge	16
Total			**1451**

Overview of the three different segments (Rule Extraction, Rule Comprehension, and Rule Compliance) and six subsets (Retrieval, Compilation, Definition, Presence, Dimension, and Functional Performance) in DesignQA.

Goal & Strategy

This hackathon challenge invites participants to develop a method that outperforms the current top-performing model (GPT-4o-AllRules) on DesignQA. Possible approaches include:

- Utilizing or tuning retrieval-augmented generation (**RAG**). Various VLMs have been coupled with off-the-shelf LlamaIndex RAG and tested on DesignQA. It was found that the off-the-shelf RAG failed to provide the models with the appropriate excerpts of engineering documentation. Improving or tuning a RAG system could improve models' scores on DesignQA.
- **Fine-tuning** an LLM or VLM. DesignQA's question-answer pairs could be used to generate larger quantities of **synthetic data**, which could in turn be used to fine-tune a model to perform better the tasks in DesignQA.
- **Prompt engineering** methods. Perhaps appending specific instructions ahead of the questions or changing a VLM's system prompt significantly improves the model's score on DesignQA.

Benchmark Details

The benchmark is designed to reflect the complexity and realism of engineering design challenges, drawing from the **140-page Formula SAE Rule document**, a real-world guide used by student teams to build functional race vehicles. It is distinguished by:

1. **Authentic Source Material**: Built on actual engineering rules and practices, similar in structure to professional standards like the **Formula 1 Technical Regulations** and NASA technical documents.

2. **Real-World Data**: Incorporates genuine **CAD models and test data** from the MIT Motorsports team, offering a depth of engineering context unavailable in most public datasets.

3. **Expert-Curated Questions**: The 1449 QAs were manually crafted and reviewed by contributors from academia, industry (Autodesk), and student teams, ensuring high quality beyond typical crowdsourced or synthetic data.

The benchmark is divided into **three segments** (each further divided into two subsets) that simulate the engineering design workflow:

- **Segment 1-Rule Extraction**: Tests the ability to locate relevant information in lengthy rule documents.

- **Segment 2-Rule Comprehension**: Evaluates recognition of technical terminology within visual designs.

- **Segment 3-Rule Compliance**: Assesses understanding of whether a design or data conforms to the specified rules.

Keep in mind that the number of QA pairs in each benchmark segment is not constant!

Submission

- You may not use any of the DesignQA benchmark data, or the original PDF, to explicitly train a model, as the whole benchmark is considered the 'test set'. You may, however, use DesignQA data as a seed for synthetic data generation.
- Please submit 6 text files as an Issue to our Github. There should be one text file for each subset of the benchmark, which contains your approach's score on that subset.
- Provide a Github repo link with your group's approach, so that the text file scores you provide could be reproduced if necessary.
- Also, please include the final presentation slides in the Github repo.

Judgment Criteria

Category	Criteria	Score
Quantitative evaluation (40%)	• Teams will report their evaluation metrics in a `results.txt` file. • Metrics will be reported according to the benchmark rules found here.	> 90% (9-10 pts) 70-80% (7-8 pts) 50-60% (5-6 pts) 30-40% (3-4 pts) < 30% (1-2 pts)
Qualitative evaluation (40%)	• Teams will present an overview of their method. • Scientific soundness of the approach. • Readiness of the idea and the approach. • Evaluation of future direction or proposed work given more time.	Excellent (9-10 pts) Very good (7-8 pts) Good (5-6 pts) Limited (3-4 pts) Poor (1-2 pts)

Overall presentation (20%)	• Title, headings, labels: appropriate size, location, spelling, and content. • Demonstration of teamwork. • Structure and clarity. • Broader impact of the idea on ME subfields.	Excellent (9-10 pts) Very good (7-8 pts) Good (5-6 pts) Limited (3-4 pts) Poor (1-2 pts)

Domain Experts and Support

Annie Doris, Decode Lab, MIT

Daniele Grandi, Autodesk