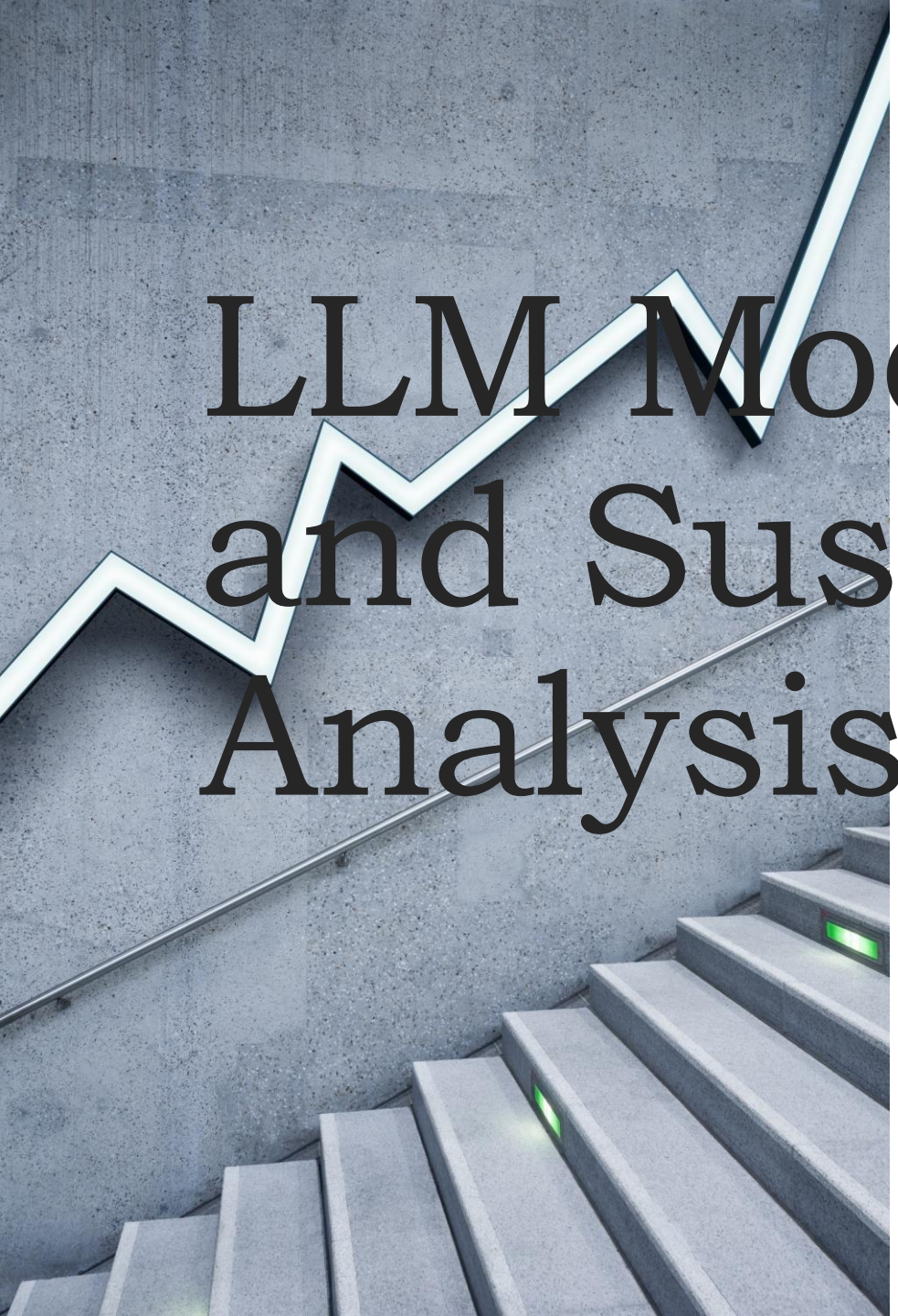
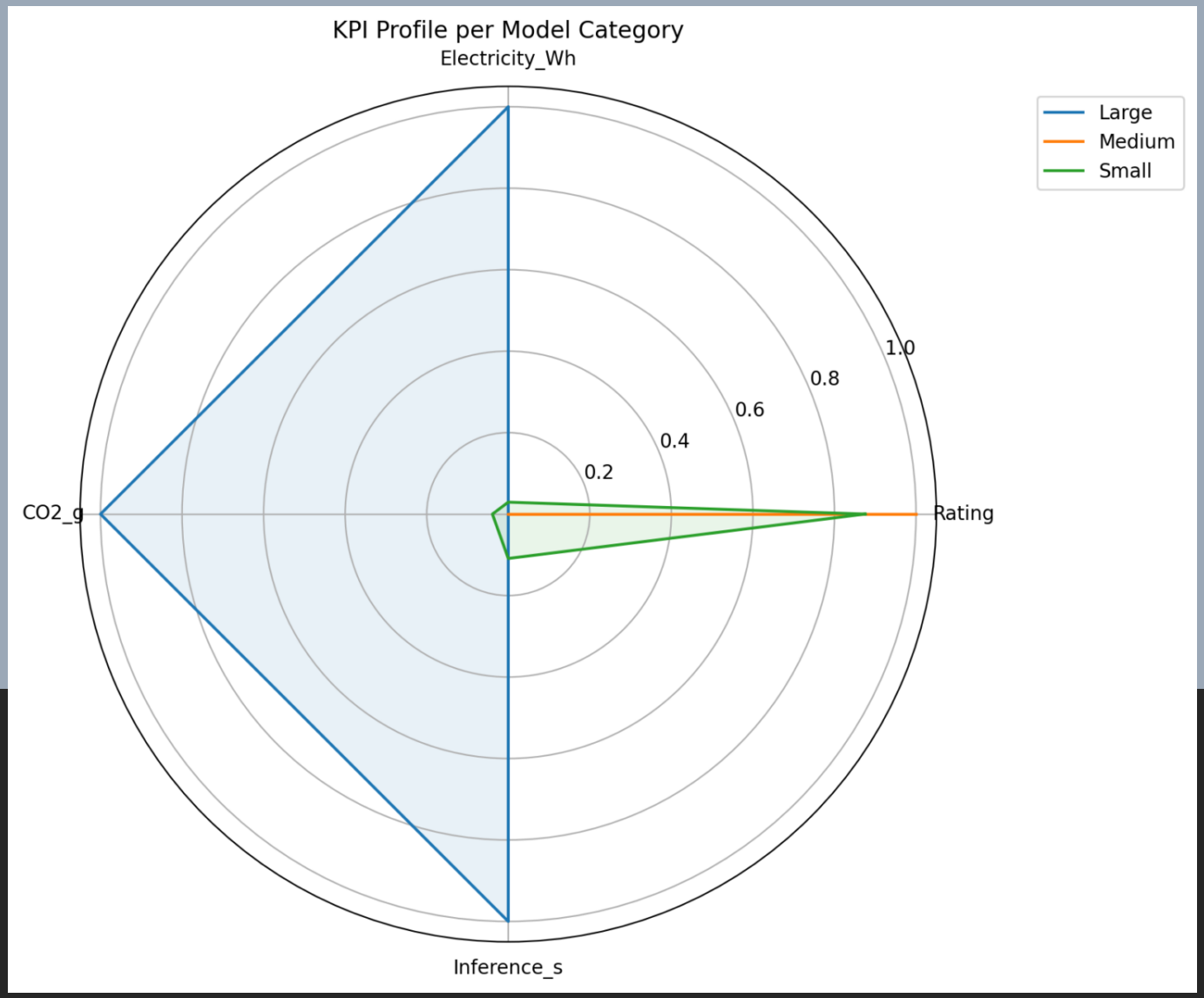


LLM Models Efficiency and Sustainability Analysis

CONTRIBUTORS:

- GAURAV CHUGH
- BIRAME MBOUP

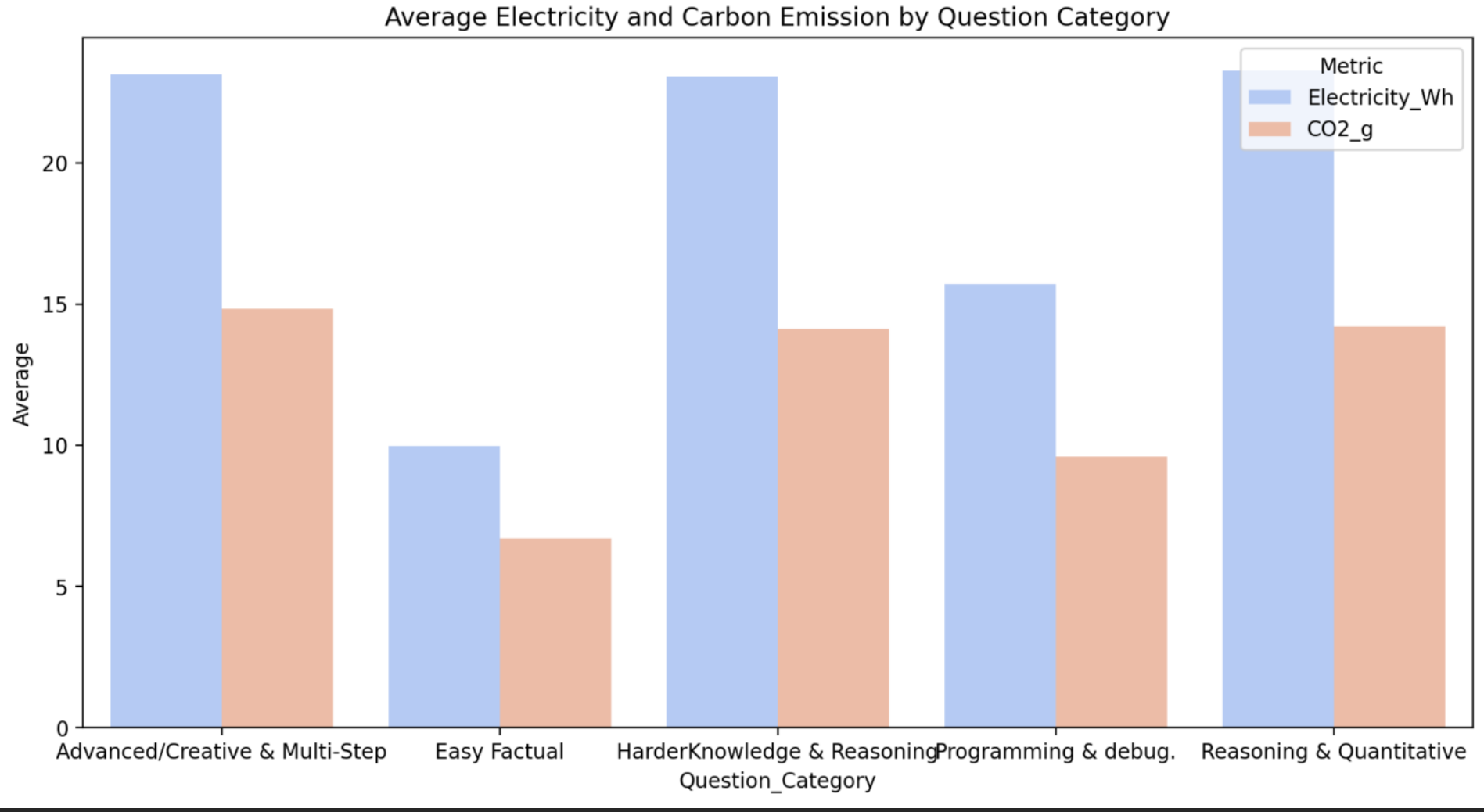
KPI Profile per Model Category

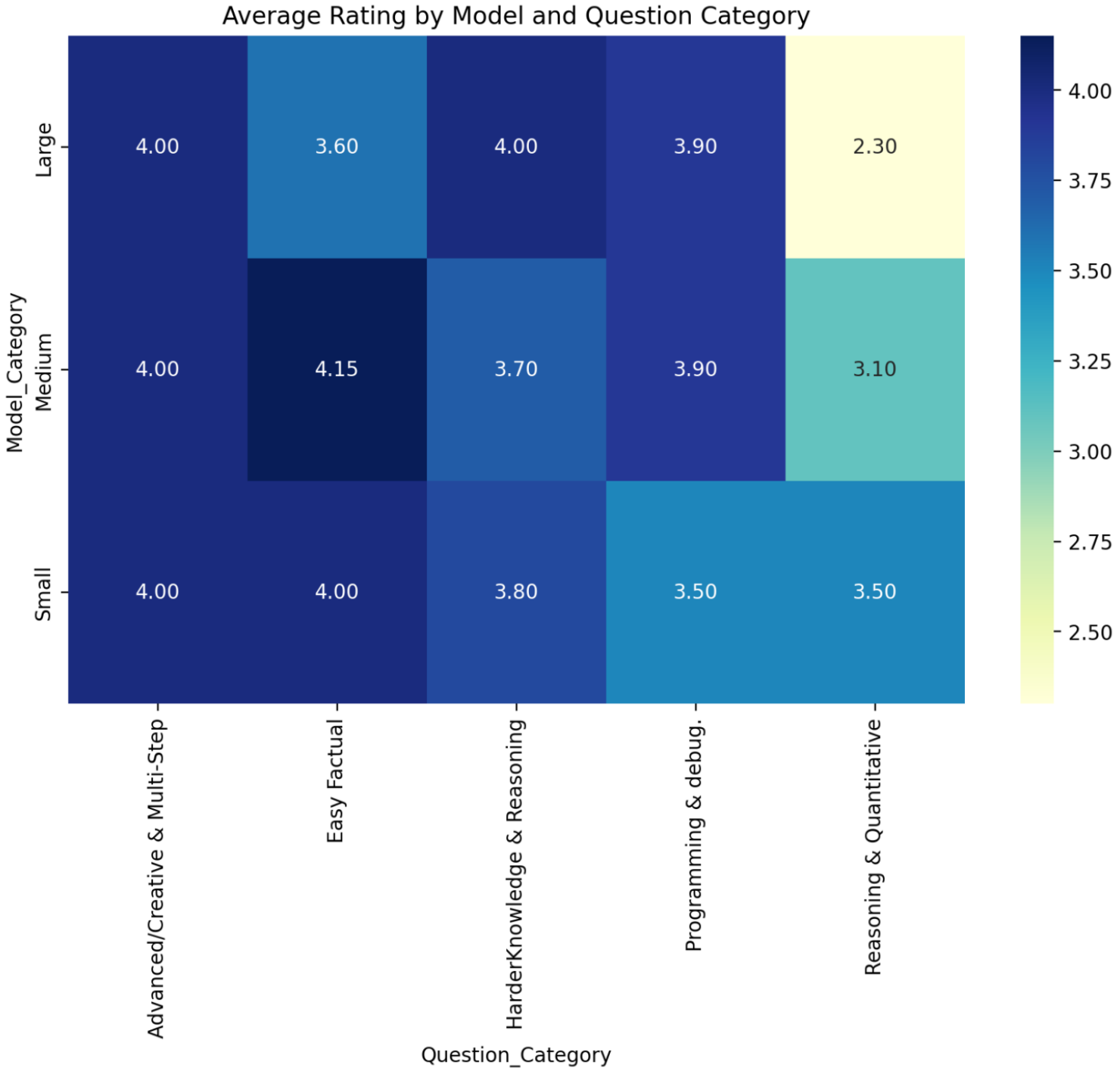


THIS RADAR CHART COMPARES MODEL CATEGORIES ACROSS KEY PERFORMANCE METRICS.

IT SHOWS THAT LARGE MODELS CONSUME THE MOST ENERGY, EMIT THE MOST CO_2, AND HAVE THE LONGEST LATENCY, WHILE MEDIUM AND SMALL MODELS MAINTAIN STRONG RATINGS WITH FAR BETTER EFFICIENCY.

- This chart compares the **average electricity consumption (Wh)** and **CO_2 emissions (g)** across different **question categories.**
  It shows that **complex tasks**—like *Advanced/Creative, Harder Knowledge, and Reasoning & Quantitative questions*—consume significantly more energy and produce higher carbon emissions than *Easy Factual* ones.



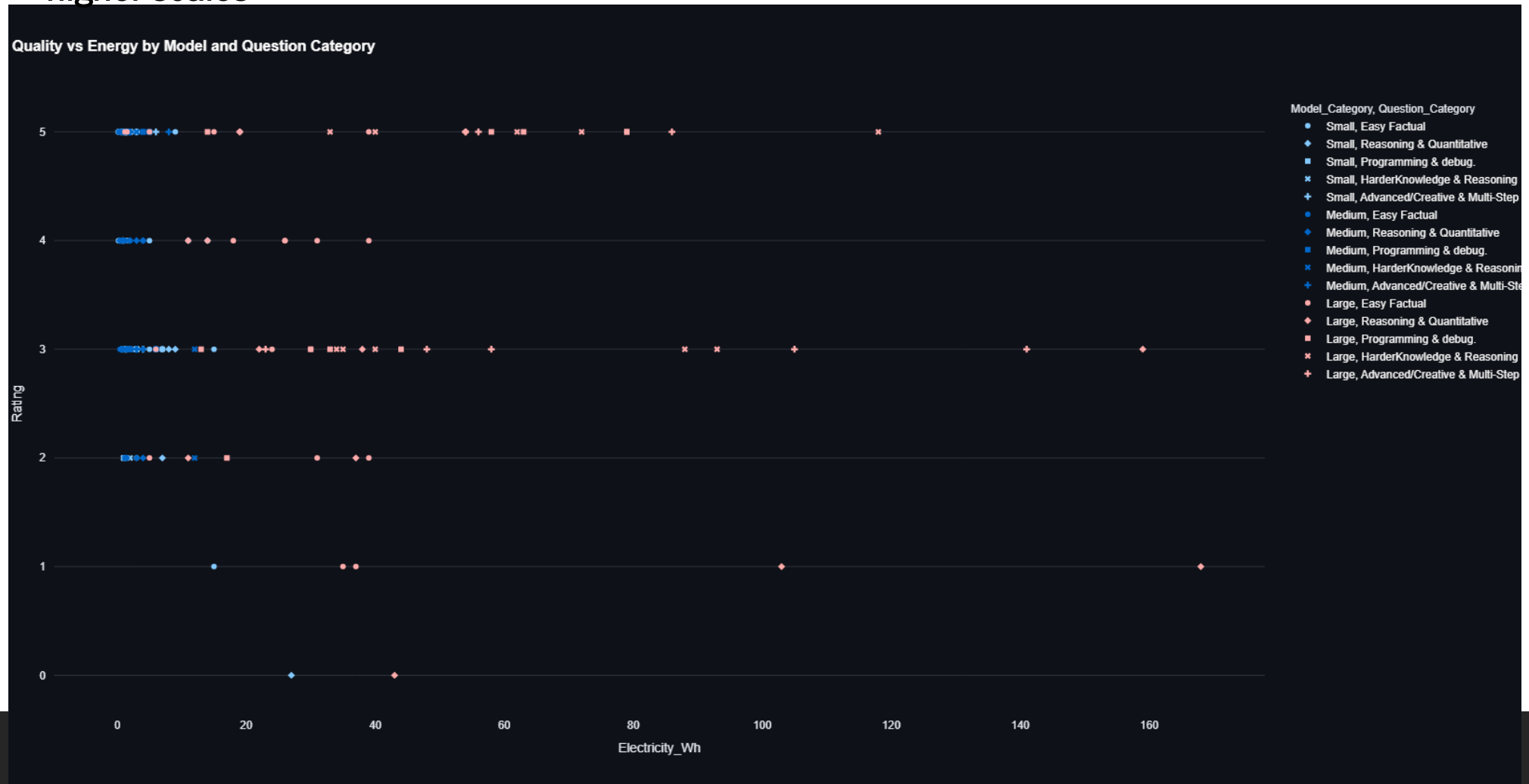


Average Rating by Model vs Question Category

- ❑ Math and reasoning tasks show higher energy and latency costs.
- ❑ Factual and descriptive questions are handled efficiently across small/medium models.

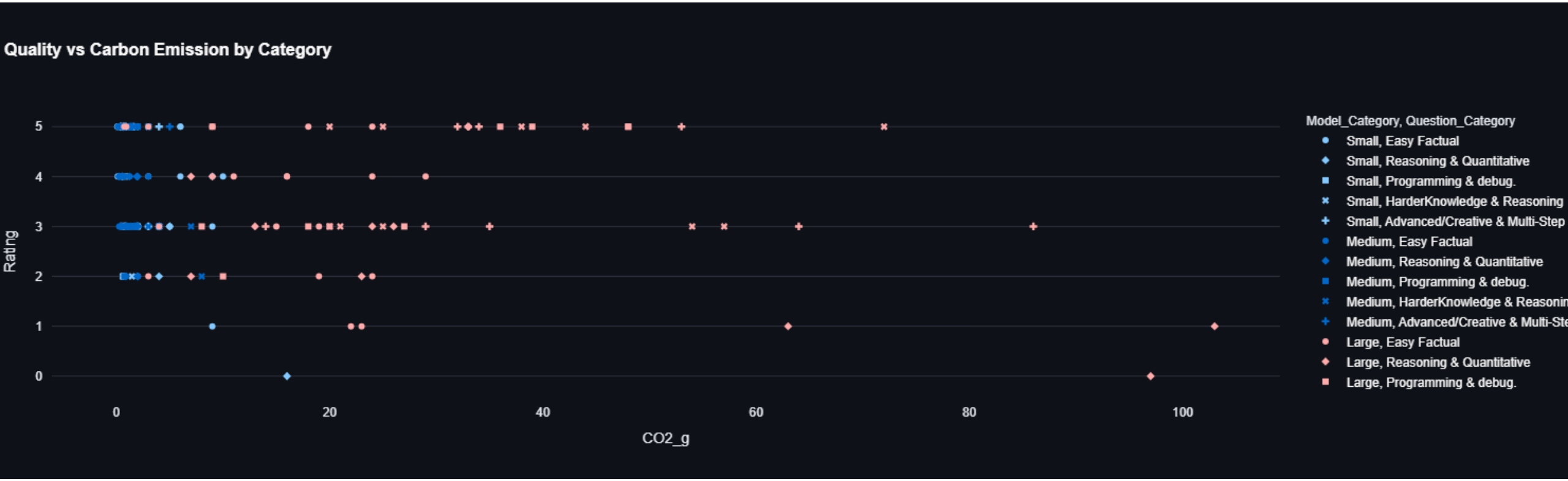
Quality vs Energy Consumption (wh)

This scatter plot shows the **relationship between model quality (Rating)** and **electricity consumption (Wh)** across **different model sizes and question categories.** It reveals that **small and medium models** achieve **high ratings (4–5)** with **low energy use**, while **large models** consume significantly more energy for similar or lower quality outputs—indicating **reduced efficiency at higher scales.**



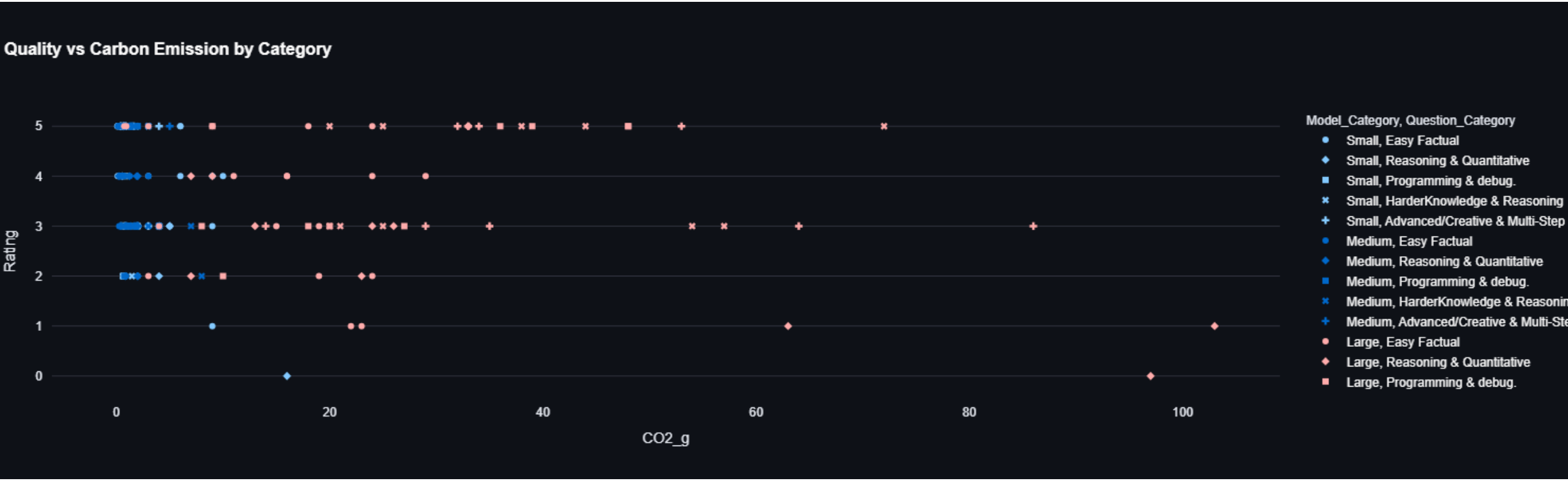
📈 Quality vs Carbon Emission (Cost Proxy)

This chart shows the relationship between model quality and carbon emissions across model sizes. **Small and medium models** deliver **high-quality results with lower CO_2 emissions**, while **large models** emit much more for similar quality.



Latency Distribution by Model Category

This chart compares model quality with carbon emissions across different model and question types. It shows that **smaller and medium models maintain high ratings with minimal CO_2 output**, while **larger models generate much higher emissions for similar performance.**



Best Trade-Off Scores

Rank	Model	Score	Remarks
1	Gemma 3nb	7.00	Best overall balance
2	GPT 20 OSS	5.23	Strong trade-off model
3	Mistral Small	4.71	Efficient but moderate performance
4–6	Llama 3.1 / GPT-5 / DeepSeek R1	Below 0	High cost, poor efficiency

This table ranks models based on a combined **trade-off score** of quality, energy use, CO_2 emissions, and latency. **Gemma 3nb** achieved the **best overall balance**, while **large models like GPT-5 and DeepSeek R1** scored lowest due to high energy and carbon costs.

Conclusion & Recommendations

- Best Efficiency: Gemma 3nb and GPT 20 OSS
- High Carbon/Energy Models: GPT-5 and DeepSeek R1
- Trade-off Strategy: Favor medium/small models for sustainable AI inference.
- Future Work:
  - Include training energy analysis
  - Explore hardware-level optimizations
  - Integrate green AI metrics in model benchmarking