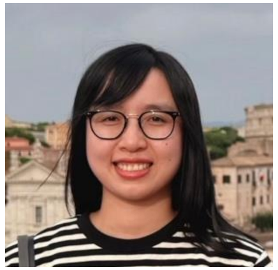
FailureSensorIQ:

A Multi-Choice QA Dataset for Understanding Sensor Relationships and Failure Modes



Authors:

Christodoulos Constantinides[2]* Dhaval Patel[1]* Shuxin Lin[1]*
Claudio Guerrero[2] Sunil Dagajirao Patil[2] Jayant Kalagnanam[1]

[1]IBM TJ Watson Research Center [2]IBM
*Equal contribution

IBM Research

Background : Industrial Assets

2 Personas

Data Scientist

Reliability Engineer

10 Industrial Assets

...

55 Failure Modes

...

53 Sensors

...

Tasks: Failure Mode and Effect Analysis

IoT sensors → capture signals: temperature, vibration, power, etc.

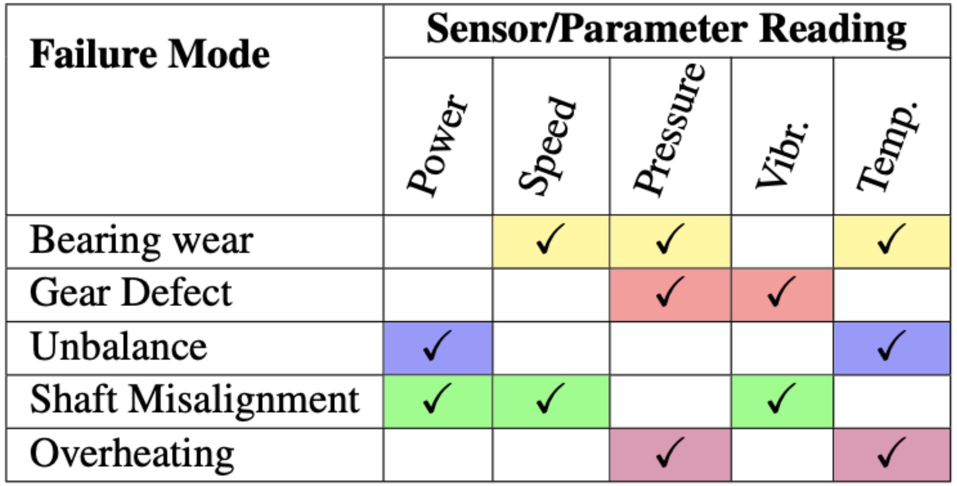
Failure Mode and Effects Analysis (FMEA) links failures ↔ sensors.

Types of queries based on query direction:

- Failure Mode to Sensors (FM2Sensor)
- Sensor to Failure Mode (Sensor2FM)

Types of queries based on logical reasoning:

- **Selection**: Identify relevant items (✓ present)
- **Elimination**: Identify irrelevant items (✓ absent)



Failure Mode	**Sensor/Parameter Reading**				
	Power	Speed	Pressure	Vibr.	Temp.
Bearing wear		✓	✓		✓
Gear Defect			✓	✓	
Unbalance	✓				✓
Shaft Misalignment	✓	✓		✓	
Overheating			✓		✓

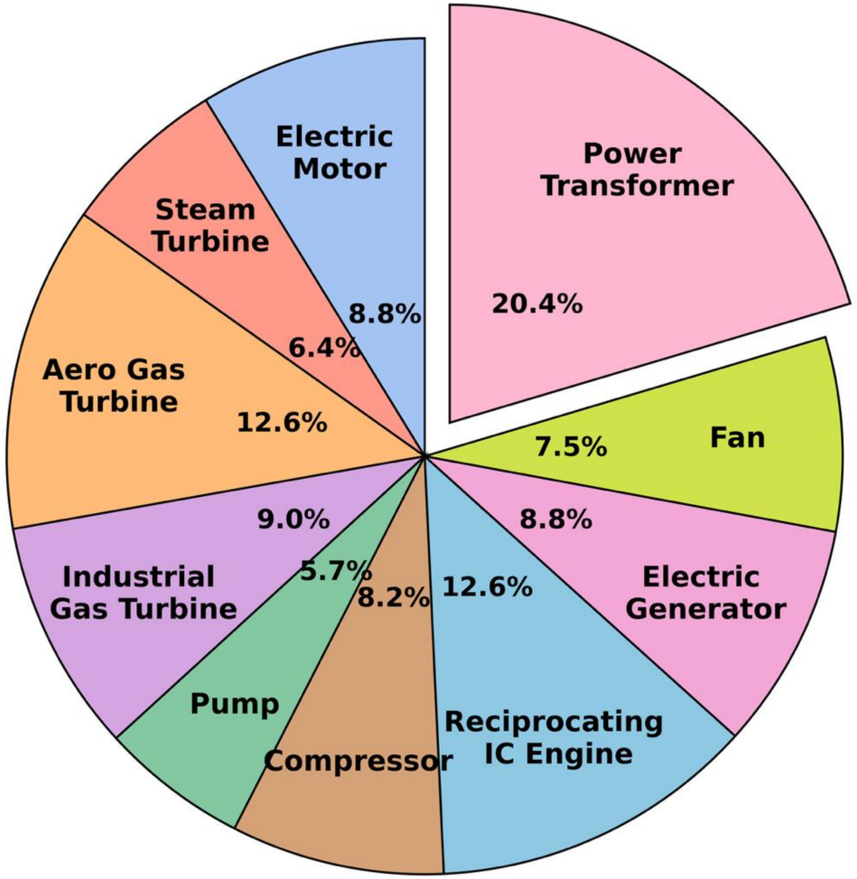
Table 1: Expert Knowledge: Failure Faults ↔ Sensors/Parameters: ✓indicates that parameter or sensor change if failure occurs

Dataset Overview

FailureSensorIQ, is a multiple-choice QA dataset that explores the relationships between sensors and failure modes for 10 industrial asset. The dataset consists of 8,296 questions with

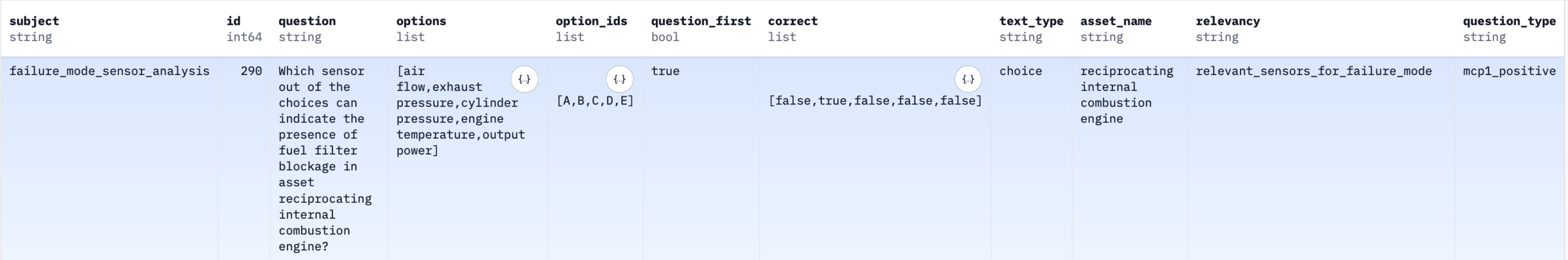
- 2,667 single-correct-answer questions (SC-MCQA)
- 5,629 multi-correct-answer questions (MC-MCQA)

The dataset leverages the information found in ISO Standards documents and expert crafted question templates guaranteeing credibility.



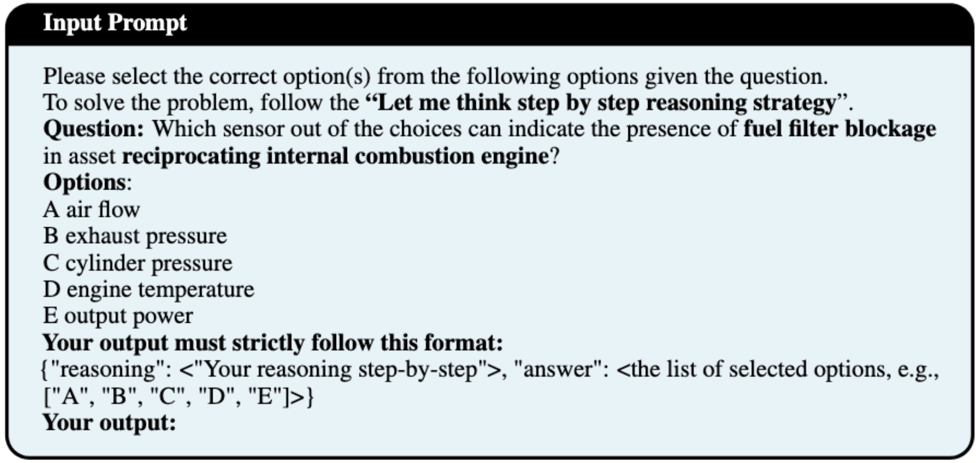
IBM Research

FailureSensorIQ: Example Question



subject string	id int64	question string	options list	option_ids list	question_first bool	correct list	text_type string	asset_name string	relevancy string	question_type string
failure_mode_sensor_analysis	290	Which sensor out of the choices can indicate the presence of fuel filter blockage in asset reciprocating internal combustion engine?	[air flow,exhaust pressure,cylinder pressure,engine temperature,output power]	[A,B,C,D,E]	true	[false,true,false,false,false]	choice	reciprocating internal combustion engine	relevant_sensors_for_failure_mode	mcp1_positive

Dataset View in HuggingFace (ibm-research/FailureSensorIQ)



Input Prompt

Please select the correct option(s) from the following options given the question.
To solve the problem, follow the **"Let me think step by step reasoning strategy**".
Question: Which sensor out of the choices can indicate the presence of **fuel filter blockage** in asset **reciprocating internal combustion engine**?
Options:
A air flow
B exhaust pressure
C cylinder pressure
D engine temperature
E output power
Your output must strictly follow this format:
{"reasoning": <"Your reasoning step-by-step">, "answer": <the list of selected options, e.g., ["A", "B", "C", "D", "E"]>}
Your output:

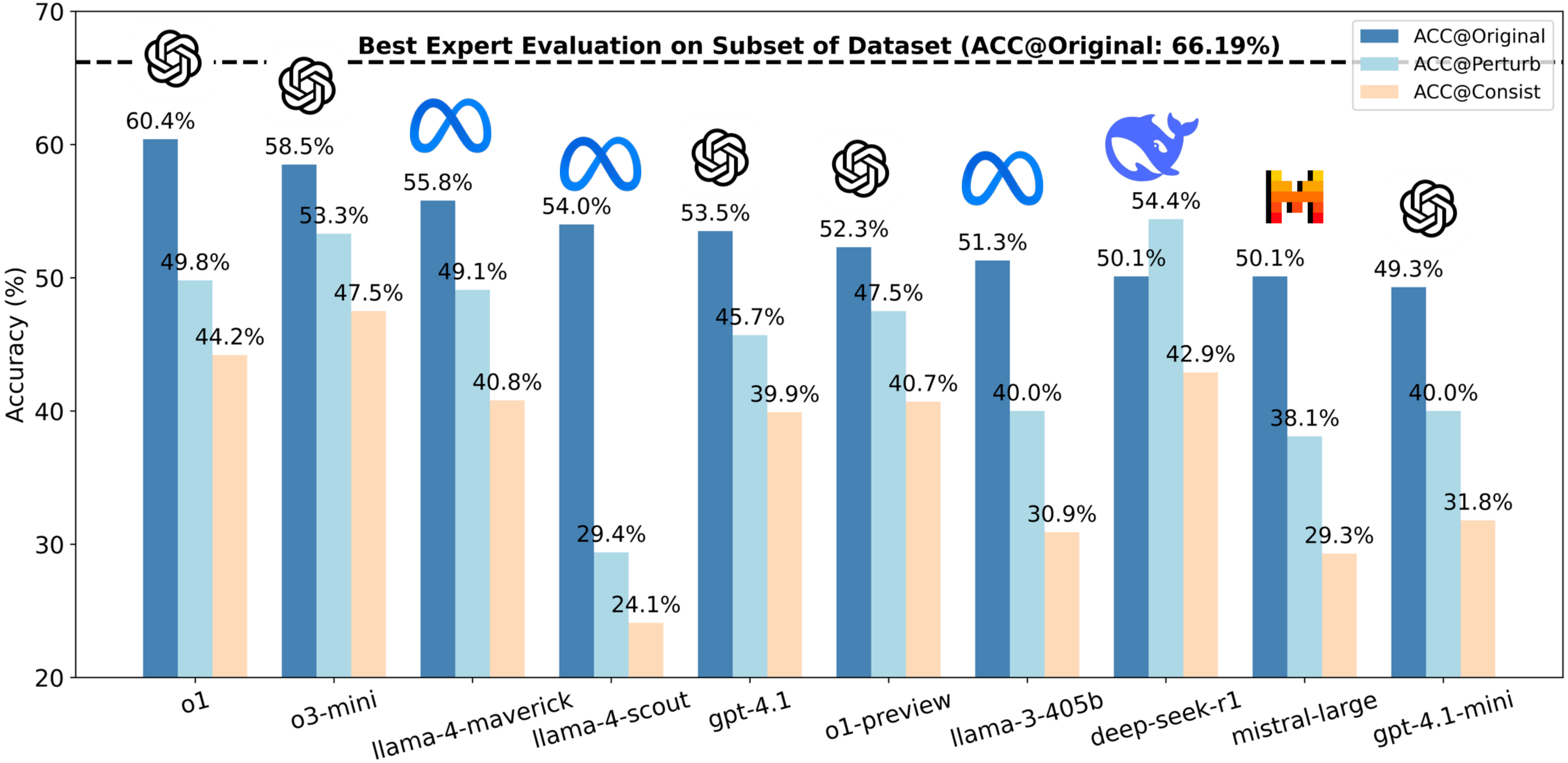
Input prompt constructed from the data

IBM **Research**

LLM Evaluation on FailureSensorIQ

- A comprehensive evaluation pipeline
  - Perturbation
  - Uncertainty
  - Complexity
- 6 Evaluation Metrics
  - Accuracy (Acc@Original)
  - Perturbed Accuracy (Acc@Perturb)
  - Consistency-Based Accuracy (Acc@Consist)
  - Set Size (SS)
  - Coverage Rate (CR)
  - Uncertainty-Adjusted Accuracy (UAcc)

Performance of SC-MCQA



IBM **Research**

Performance of MC-MCQA

Overall performance suggests that exact selection of multiple true answers remains a difficult task, especially without explicit guidance on how many options are correct.

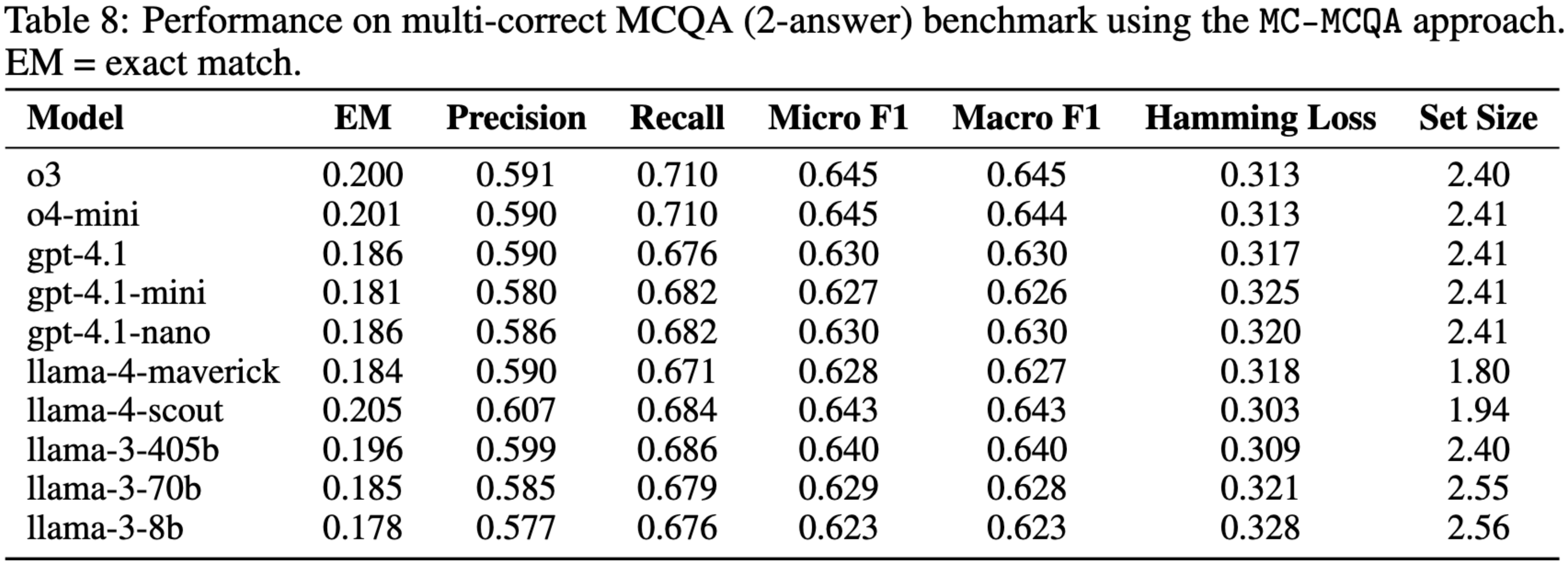


Table 8: Performance on multi-correct MCQA (2-answer) benchmark using the MC-MCQA approach. EM = exact match.

Model	EM	Precision	Recall	Micro F1	Macro F1	Hamming Loss	Set Size
o3	0.200	0.591	0.710	0.645	0.645	0.313	2.40
o4-mini	0.201	0.590	0.710	0.645	0.644	0.313	2.41
gpt-4.1	0.186	0.590	0.676	0.630	0.630	0.317	2.41
gpt-4.1-mini	0.181	0.580	0.682	0.627	0.626	0.325	2.41
gpt-4.1-nano	0.186	0.586	0.682	0.630	0.630	0.320	2.41
llama-4-maverick	0.184	0.590	0.671	0.628	0.627	0.318	1.80
llama-4-scout	0.205	0.607	0.684	0.643	0.643	0.303	1.94
llama-3-405b	0.196	0.599	0.686	0.640	0.640	0.309	2.40
llama-3-70b	0.185	0.585	0.679	0.629	0.628	0.321	2.55
llama-3-8b	0.178	0.577	0.676	0.623	0.623	0.328	2.56

Other Experiments Conducted

- Impact of Reasoning-Based Prompting
- AI Agent with External Knowledgebase
- Human Evaluation
- LLMFeatureSelect: scikit-learn Transformer

Links

- HuggingFace Dataset: https://huggingface.co/datasets/ibm-research/FailureSensorIQ
- GitHub Repository: https://github.com/IBM/FailureSensorIQ
- Arxiv Paper: https://arxiv.org/abs/2506.03278

HuggingFace

GitHub

Arxiv