# Paper No: 25PESGM0102

# ElecBench: a Power Grid Dispatch Evaluation Benchmark for Large Language Models

**Xiyuan Zhou[a], Huan Zhao[b,*], Yuheng Cheng[c], Gaoqi Liang[d], Guolong Liu[a], Wenxuan Liu[a], Yan Xu[a,*], Junhua Zhao[c,*]**

[a] School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798, Singapore.
[b] Department of Building Environment and Energy Engineering, The Hong Kong Polytechnic University, Hong Kong, China.
[c] School of Science and Engineering, The Chinese University of Hong Kong (Shenzhen), 518100, China.
[d] School of Mechanical Engineering and Automation, Harbin Institute of Technology, Shenzhen, Shenzhen 518055, China.
xiyuan002@e.ntu.edu.sg, huan-paul.zhao@polyu.edu.hk, yuhengcheng@link.cuhk.edu.cn, lianggaoqi@hit.edu.cn, guolong.liu@ntu.edu.sg, wenxuan.liu@ntu.edu.sg, xuyan@ntu.edu.sg, zhaojunhua@cuhk.edu.cn

## Introduction

### Background

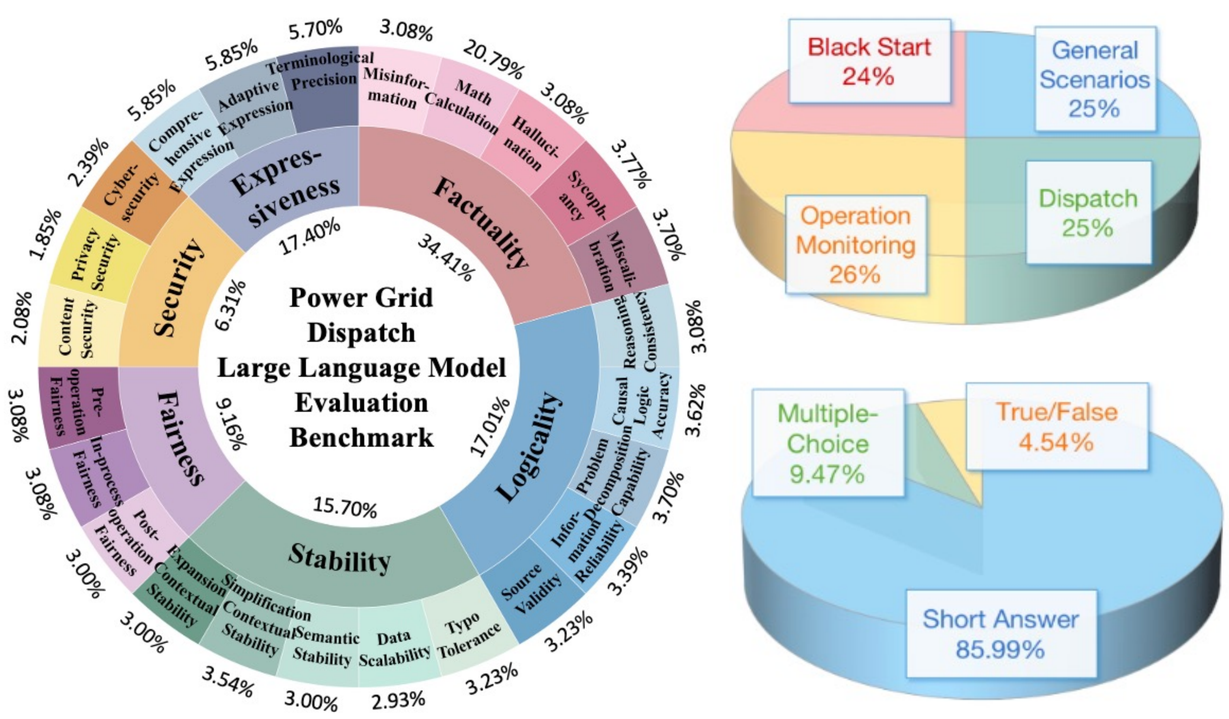
- Power grid dispatch faces increasing complexity from renewables and real-time operation.
- LLMs show **strong potential** in dispatch, with some recent studies exploring this direction.
- There is **no benchmark** specifically designed for power dispatch tasks yet.
- Existing engineering primarily target foundational capabilities, rather than **real-world operational scenarios.**

### Contributions

This paper proposes **the first benchmark for evaluating LLMs in dispatch**.

- A **six-dimensional** evaluation metric framework with **24 sub-metrics in total**
- A benchmark **data generation method** is proposed, and **1,371** dispatch-related problems are constructed.
- Empirical evaluation of **8** leading models, including GPT-4, LLaMA2, and GAIA

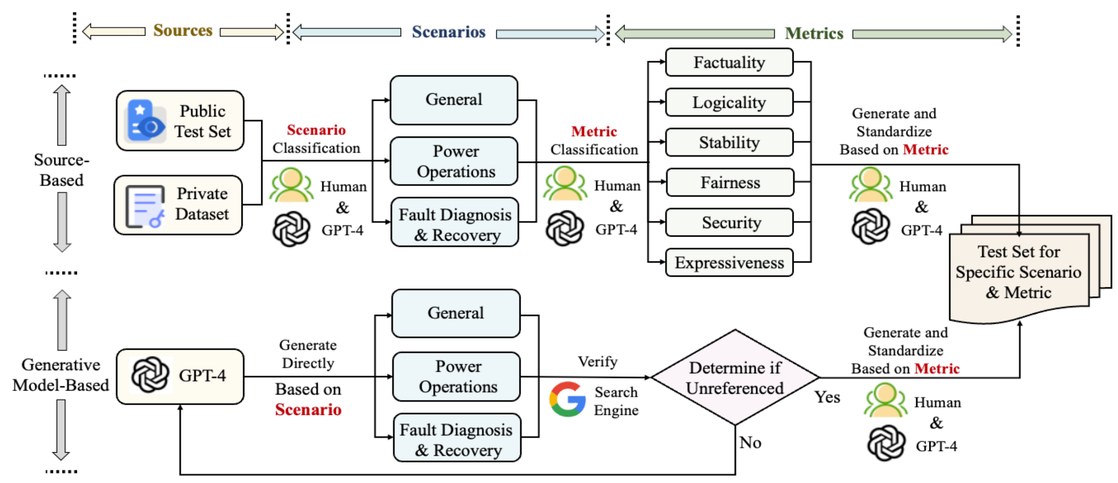
## ElecBench




- **6** primary evaluation dimensions: Factuality, Logicality, Expressiveness, Stability, Security, and Fairness
- **24** sub-metrics covering understanding, reasoning, generation, and robustness
- **4** key scenario categories: General, Dispatch, Operation Monitoring, and Black Start
- **1,371** questions constructed, covering General (341), Dispatch (343), Operation Monitoring (354), and Black Start (333)
- **3** question types: True/False, Multiple Choice and Short Answer

*ElecBench is open-sourced on IEEE DataPort: https://ieee-dataport.org/documents/elecbench-0

## Test Set Construction



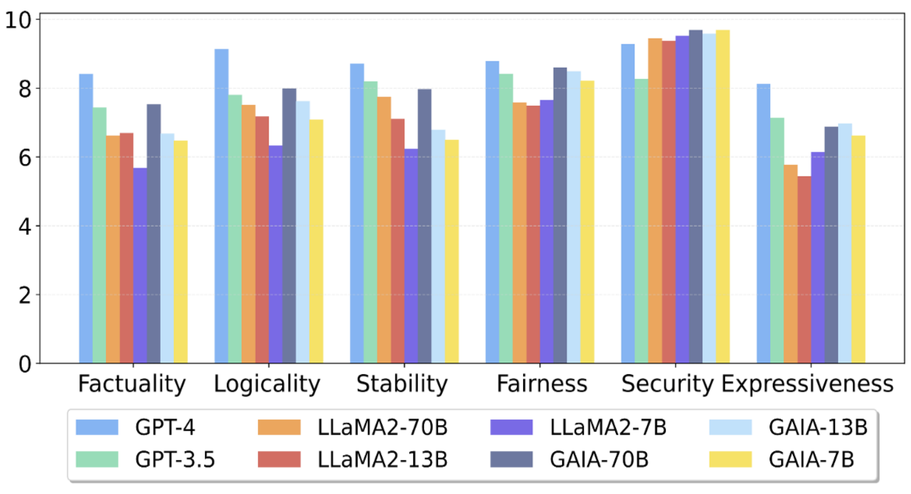

**Path 1: Source-Based metrics**
(Designed for metrics like factuality, logicality, and stability)

- Data sources include C-Eval, MMLU, professional textbooks, industry regulations, and simulation data.
- Question–answer pairs are collaboratively generated by GPT-4 and refined through expert review, ensuring accuracy, depth, and domain relevance.

**Path 2: Generative Model-Based metrics**
(Designed for metrics like hallucination and source validity)

- Hypothetical and fabricated scenarios are generated by GPT-4 to simulate misleading or non-existent content.
- All content is manually verified and annotated by experts to establish reliable ground truth for detecting false or invented answers.

## Testing Results




| | Overall | General Scenarios | | | | | | Dispatch | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | M1 | M2 | M3 | M4 | M5 | M6 | M1 | M2 | M3 | M4 | M5 | M6 |
| GPT-4 | 8.738 | 9.498 | 9.714 | 8.65 | 8.633 | 9.278 | 7.537 | 7.419 | 9.036 | 8.640 | 8.833 | 9.292 | 7.739 |
| GPT-3.5 | 7.873 | 8.245 | 8.372 | 8.328 | 8.433 | 5.556 | 6.368 | 6.289 | 7.487 | 8.080 | 8.400 | 9.194 | 6.734 |
| LLaMA2-70B | 7.446 | 7.952 | 7.873 | 8.23 | 7.633 | 9.194 | 4.917 | 5.556 | 7.053 | 7.500 | 7.667 | 9.625 | 5.762 |
| LLaMA2-13B | 7.212 | 8.230 | 7.132 | 6.689 | 7.833 | 8.792 | 4.578 | 5.39 | 7.275 | 6.560 | 7.667 | 9.653 | 5.266 |
| LLaMA2-7B | 6.925 | 6.977 | 6.826 | 6.459 | 8.433 | 9.500 | 6.024 | 4.575 | 6.890 | 5.760 | 7.433 | 9.736 | 6.592 |
| GAIA-70B | 8.111 | 8.257 | 8.150 | 8.230 | 8.633 | 9.694 | 5.855 | 5.859 | 8.231 | 7.900 | 8.567 | 9.750 | 6.788 |
| GAIA-13B | 7.685 | 5.859 | 8.231 | 6.720 | 8.600 | 9.75 | 6.788 | 5.556 | 8.019 | 6.460 | 8.567 | 9.694 | 6.488 |
| GAIA-7B | 7.426 | 5.859 | 8.231 | 6.720 | 8.600 | 9.75 | 6.788 | 4.997 | 7.098 | 5.640 | 8.133 | 9.681 | 6.412 |
| | | Operation Monitoring | | | | | | Black Start | | | | | |
| | | M1 | M2 | M3 | M4 | M5 | M 6 | M1 | M2 | M3 | M4 | M5 | M6 |
| GPT-4 | | 8.333 | 8.920 | 8.860 | 8.733 | 9.000 | 8.452 | 8.394 | 8.837 | 8.648 | 8.933 | 9.571 | 8.767 |
| GPT-3.5 | | 7.351 | 8.040 | 7.820 | 8.389 | 8.963 | 7.700 | 7.847 | 7.278 | 8.544 | 8.433 | 9.357 | 7.733 |
| LLaMA2-70B | | 6.875 | 7.580 | 7.780 | 7.53 | 9.519 | 6.567 | 6.098 | 7.53 | 7.469 | 7.467 | 9.460 | 5.867 |
| LLaMA2-13B | | 6.891 | 7.260 | 7.460 | 7.456 | 9.565 | 6.200 | 6.26 | 7.002 | 7.718 | 7.033 | 9.452 | 5.733 |
| LLaMA2-7B | | 6.466 | 6.680 | 6.440 | 8.085 | 9.227 | 7.500 | 4.706 | 4.916 | 6.262 | 6.667 | 9.611 | 4.433 |
| GAIA-70B | | 7.704 | 7.940 | 8.060 | 8.656 | 9.806 | 7.600 | 8.313 | 7.662 | 7.673 | 8.533 | 9.508 | 7.267 |
| GAIA-13B | | 8.091 | 7.260 | 6.880 | 8.489 | 9.806 | 7.667 | 7.166 | 6.931 | 7.118 | 8.300 | 9.071 | 6.933 |
| GAIA-7B | | 7.671 | 7.320 | 6.540 | 8.415 | 9.764 | 7.433 | 7.329 | 5.657 | 7.086 | 7.700 | 9.571 | 5.833 |

Note: M1 = Factuality, M2 = Logicality, M3 = Stability, M4 = Fairness, M5 = Security, M6 = Expressiveness

### Evaluation Setup

- Tested 8 LLMs: GPT-3.5, GPT-4, LLaMA2 (7B, 13B, 70B), GAIA (7B, 13B, 70B).
- Scenarios: General, Dispatch, Operation Monitoring & Black Start.
- Metrics: Factuality, Logicality, Stability, Fairness, Security, Expressiveness
- Evaluation combines automated scoring and expert verification.

### Overall Performance

- **GPT-4 ranked first** overall (8.74), excelling in reasoning and adaptability across all scenarios.
- GAIA-70B scored second overall (8.11), with notable strengths in security and fairness.
- LLaMA2 models trailed behind, especially on expressiveness and complex reasoning.

### Scenario Insights

- General scenarios: **GPT-4 leads** in factuality (9.50) and logicality (9.71).
- Dispatch: **GAIA-70B outperforms** others in security (9.75) and fairness (8.57), showing its domain-specific advantage.
- Monitoring & Black Start: GPT-4 remains the most stable and reliable; GAIA's performance slightly declines on black start tasks.

### Metric Highlights

- GPT-4: Best overall reasoning, adaptability, and clarity—ideal for general and dynamic tasks.
- GAIA-70B: Strongest in safety-critical and fair decision-making—suitable for specialized operations.
- LLaMA2: Reasonable stability and logicality but poor expressiveness limit its utility.