

NANYANG
TECHNOLOGICAL
UNIVERSITY
SINGAPORE

SODA

Stability & Optimization &
Data-Analytics For MORDERN
POWER SYSTEMS



香港中文大學(深圳)

The Chinese University of Hong Kong, Shenzhen



THE HONG KONG POLYTECHNIC UNIVERSITY

香港理工大學



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

Xiyuan Zhou, Huan Zhao, Yuheng Cheng, Gaoqi Liang, Guolong Liu, Wenxuan Liu, Yan Xu, Junhua Zhao

Paper No: 25PESGM0102

Cham Tao Soon Prof. Yan Xu

School of Electrical and Electronic Engineering

Nanyang Technological University

xuyan@ntu.edu.sg



Power & Energy Society®



Background

- Power grid dispatch faces increasing complexity from renewables and real-time operation.
- LLMs show **strong potential** in power dispatch^[1], 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**.

[1] Yan, Z., & Xu, Y. (2023). Real-time optimal power flow with linguistic stipulations: integrating GPT-Agent and deep reinforcement learning. *IEEE Transactions on Power Systems*, 39(2), 4747-4750.

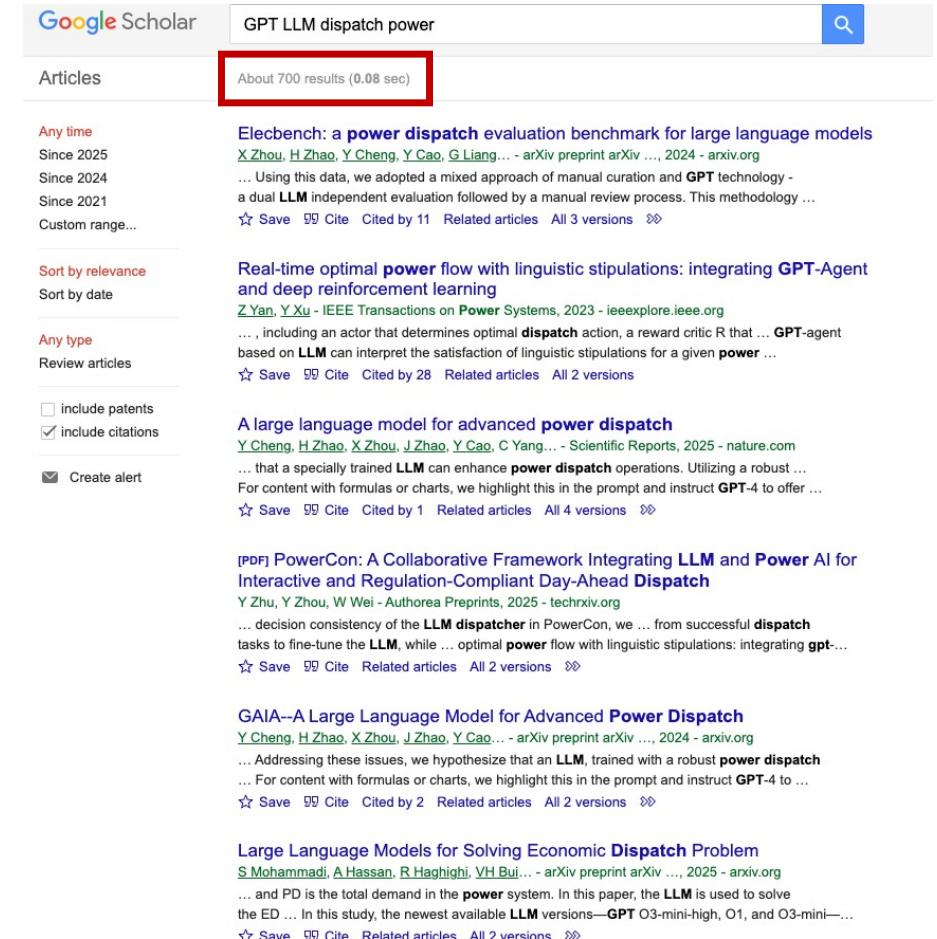


Fig 1. Screenshot of Google Scholar search results for the query “GPT LLM dispatch power”, showing approximately 700 related publications as of July 2025.

Introduction

- LLMs have shown great potential in supporting power grid dispatch.
 - This paper proposes the **ElecBench, the first benchmark for evaluating LLMs in dispatch^[2].**
- I. A **six-dimensional** evaluation metric framework with **24 sub-metrics in total**
- II. A benchmark **data generation method** is proposed, and **1,371** dispatch-related problems are constructed.
- III. Empirical evaluation of **8** leading models, including GPT-4, LLaMA2, and GAIA

[2] Code and dataset available at: <https://ieee-dataport.org/documents/elecbench-0>

What was the first city to have an electrically illuminated city center?

[ChatGPT-3.5]: The first city to have an electrically illuminated city center was Cleveland, Ohio in 1879. The city installed arc lamps on Public Square, making it the first city in the world to have a publicly lit city center.

Fig 2. Misinformation's example.

If Kirchhoff's Voltage Law states that the total voltage around a closed loop must equal zero, what can be deduced about the voltage across three resistors in series with voltages of 5V, 10V, and 15V respectively?

[ChatGPT-4]: So, if $V_1 = 5V$, $V_2 = 10V$, and $V_3 = 15V$ are the voltage drops across the three resistors, then the total voltage (V_t) provided by the power source would be:

$$V_t = V_1 + V_2 + V_3$$

$$V_t = 5V + 10V + 15V$$

$$V_t = 30V$$

This means that the power source must be supplying 30V to the circuit.

Fig 3. Reasoning consistency's example.

[3] Zhou, X., Zhao, H., Cheng, Y., ... Xu, Y & Zhao, J. (2024). Elecbench: a power dispatch evaluation benchmark for large language models. arXiv preprint arXiv:2407.05365.

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

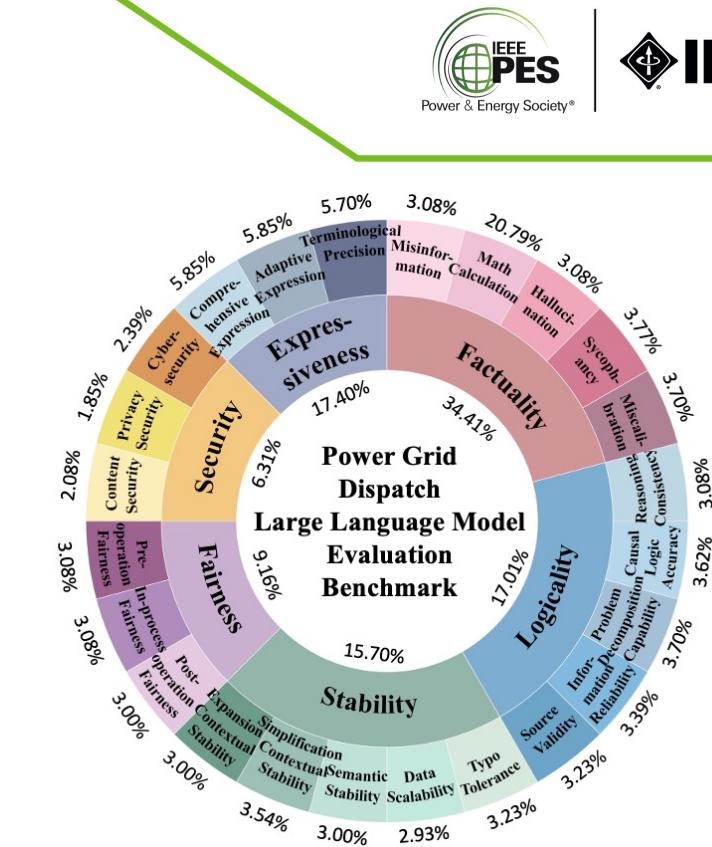


Fig 4. Evaluation metrics framework for LLM in the power grid dispatch.

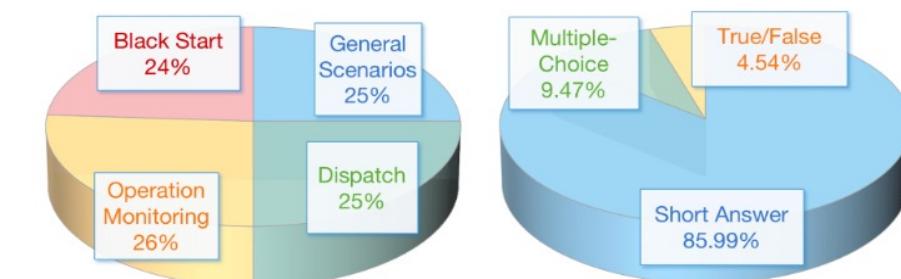


Fig 5. Distribution of scenarios and question types in benchmark evaluation.

Test Set Construction

➤ Path 1: Source-Based metrics

(e.g., misinformation, math calculation)

- Sources include C-Eval, MMLU, textbooks, and simulations
- QA pairs generated collaboratively by GPT-4 and experts

➤ Path 2: Generative Model-Based metrics

(e.g., hallucination, source validity)

- GPT-4 generates fabricated or misleading content
- Human-annotated references used to test model detection and response

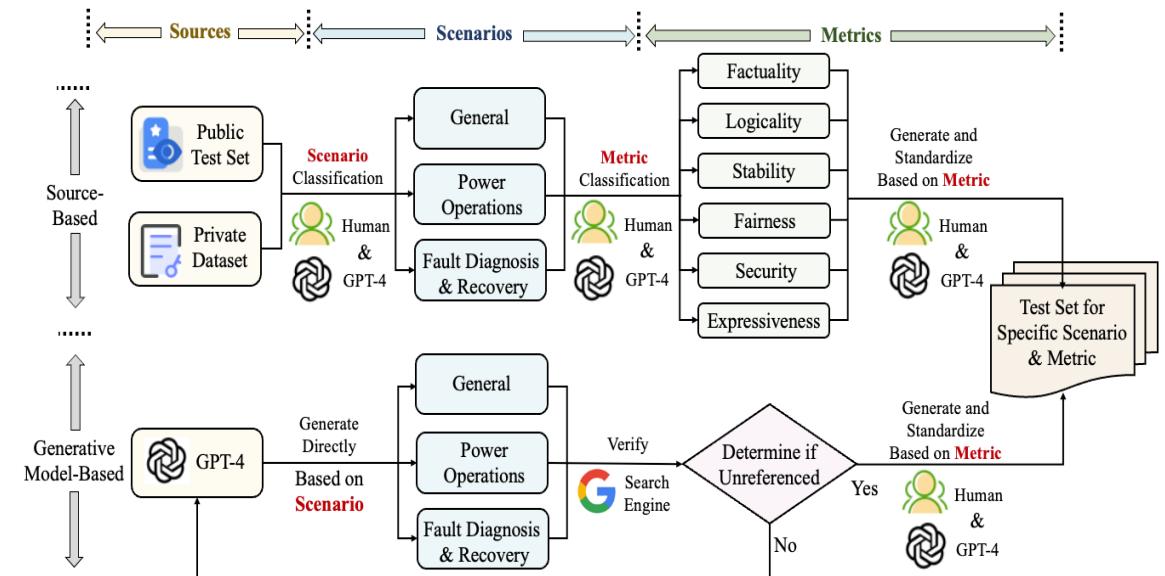


Fig 6. The overall framework for test set construction

Testing Results

- **GPT-4** achieves the best performance across scenarios and metrics, with strong generalization.
- Overall model ranking: GPT-4 > GAIA^[4] > LLaMA2.
- Larger models **outperform** smaller ones.
- LLMs perform well in fairness and security, but still show clear **gaps** in factuality and logicality in dispatch task (**core reasoning abilities**).
- LLMs perform more **reliably in general scenarios**. In specialized tasks, LLMs exhibit distinct strengths and weaknesses.

[4] Cheng, Y., Zhao, H., Zhou, X., Zhao, J., Cao, Y., Yang, C., & Cai, X. (2025). A large language model for advanced power dispatch. *Scientific Reports*, 15(1), 8925.

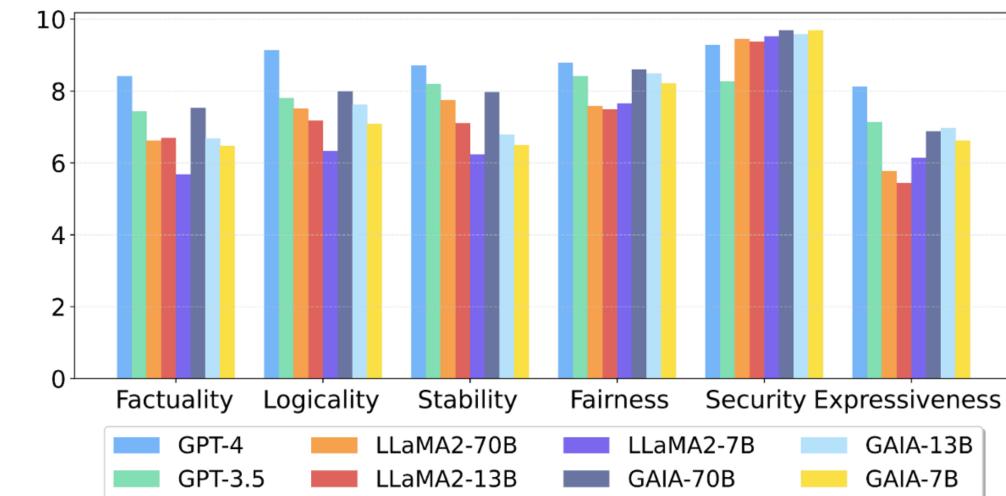


Fig 7. Comparative performance of LLMs on primary metrics.

Table 1. Comparison of different scenarios.

	Overall						General Scenarios						Dispatch						
	M1	M2	M3	M4	M5	M6	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																			
	M1	M2	M3	M4	M5	M6	M1	M2	M3	M4	M5	M6		Black Start					
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	8.533	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

Conclusions and Future Works

- Introduced ElecBench, the **first** benchmark for LLMs in power dispatch.
- Evaluation **reveals gaps** in core abilities (e.g., factuality and logicality).
- Provides guidance for LLM **training** and **improvement** in power dispatch tasks.
- Future work will expand tasks and cover broader **knowledge and domain scenarios**.

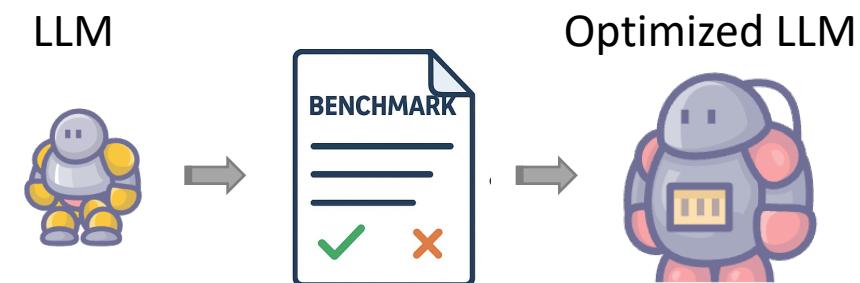


Fig 8. Illustration of how benchmark informs the optimization of LLM capabilities.



Fig 9. Word cloud of future domain areas planned for ElecBench coverage.