Benchmarking Large Language Models for the Electric Power Sector

Apurba Sakti, Ph.D.
Principal Technical Leader, EPRI

OPAI MRC Meeting
17 December 2025

https://www.epri.com/research/products/000000003002034347

www.epri.com

Motivation

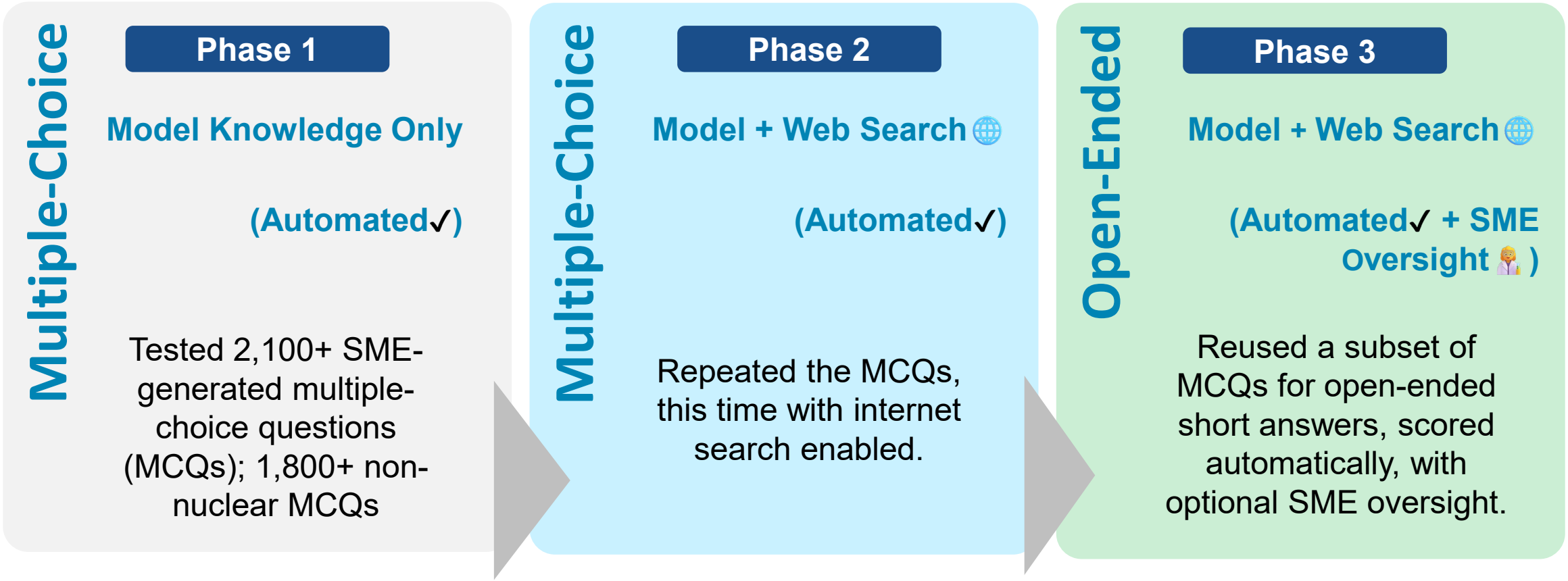
External benchmarks commonly emphasize multiple-choice formats that measure broad academic knowledge, such as math, science, and coding.

However, utilities are anticipated to derive value from AI systems on complex, open-ended queries that are focused on power-system topics and require accurate reasoning, grounding, and transparency.

This gap motivates EPRI's multi-phase approach that evaluates large language models (LLMs) on over 35 power-sector topics. This is the first step towards EPRI's evaluation of domain-augmented tools & real-world applications.

EPRI's Benchmarking Effort Progresses from Baseline Knowledge to Augmented Reasoning

Phases 1-3 leverage automated repeatable evaluations and scoring with the option of SME oversight



Automated Evaluations were Performed Using Inspect AI

Inspect AI is the UK AI Security Institute's Open-Source Framework for Evaluating LLM Performance & Reliability

MCQs

Phases 1&2

- EPRI SMEs develop 2,115 MCQs
- 1,818 of 2,115 are non-nuclear*

SAQs

Phase 3

- Subset of 399 high-quality questions selected for open-ended evaluation

Benchmark

Inspect AI used for automation

- 3 sets of runs for each phase
- Median accuracy with 95% confidence interval

AISI AI SECURITY INSTITUTE Inspect AI

inspect.aisi.org.uk

MCQs: Multiple Choice Questions, SAQs: Short Answer Questions

*Due to the sensitive nature of nuclear content and restrictions on testing LLMs with nuclear-related material, this report primarily focuses on the 1,800+ non-nuclear Q&As.

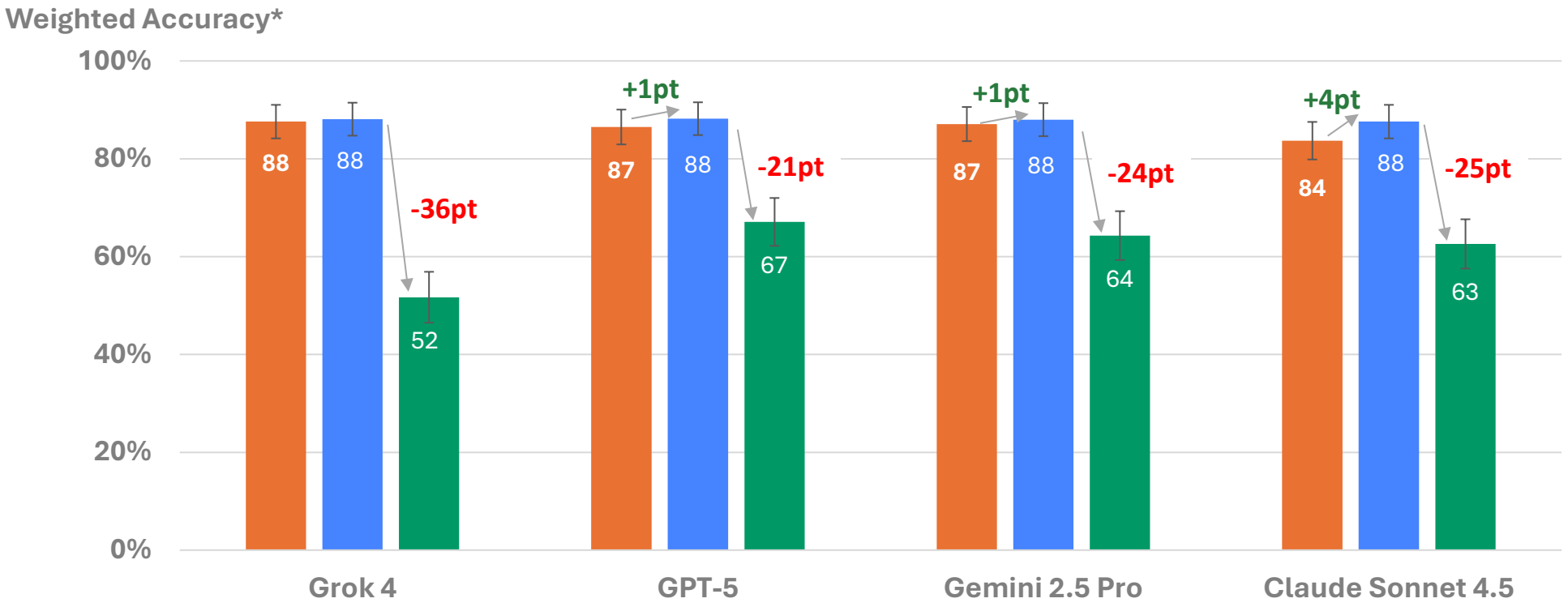
Open-Ended Questions Led to a 27pt. Accuracy Drop vs. MCQs

Some frontier LLMs answered incorrectly ~50% of the time

EPRI's Power-Systems Benchmarking – from multiple choice to open-ended questions

399 Question Subset; 95% Confidence Intervals

■ Phase 1 – MCQs (Model Knowledge) → ■ Phase 2 – MCQs (Model + Web Search) → ■ Phase 3 – Open-Ended (Model + Web Search)



Enabling web search in Phase 2 had a small effect on model accuracies. Phase 3's open-ended format proved materially harder for the LLMs

*__Difficulty-weighted__ scores reported to tighten dispersion across easy/medium/hard questions using weights of 1, 2, and 3 for easy/early-career, medium/experienced engineer, and hard/SME questions respectively. Each model was evaluated **three times** to measure run-to-run variability and ensure the results are statistically robust. **The median of the three is reported along with bars depicting the 95% confidence interval using the methodology here:** Data Analysis Toolkit 12: Weighted Averages and their Uncertainties, Adding Error Bars to Evals: A Statistical Approach to Language Model Evaluations

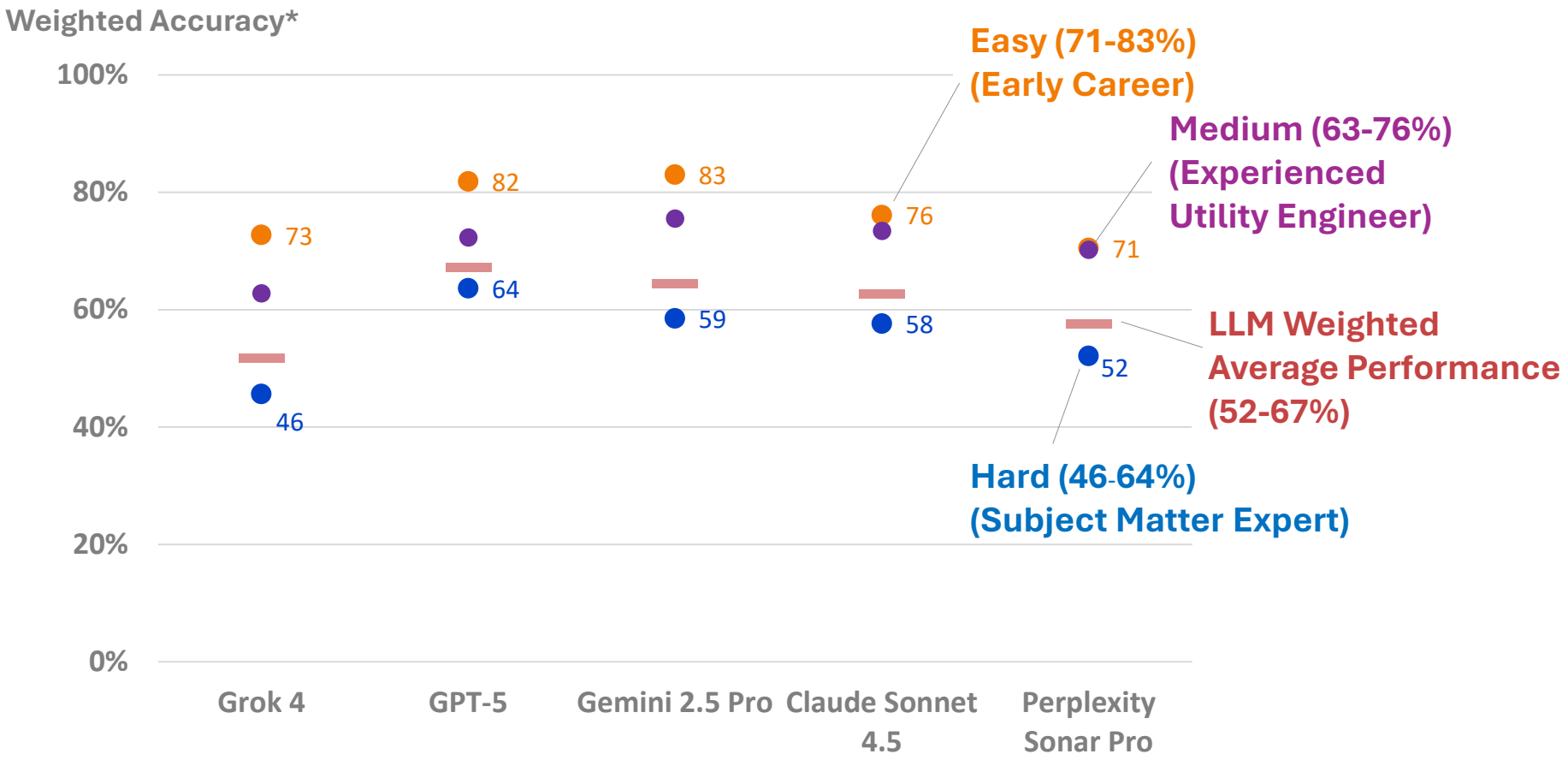
Additional LLM Details: **Grok 4:** grok/grok-4-0709; **GPT 5**: openai/gpt-5-2025-08-07; **Gemini 2.5 Pro**: google/gemini-2.5-pro; **Claude Sonnet 4.5**: anthropic/claude-sonnet-4-5-20250929

Frontier LLMs Scored 46% to 64% on Hard Open-Ended Questions

Models maintained ~71–83% accuracy on easy questions but dropped sharply on expert-level ones

Model Performance Across Easy, Medium, & Hard Questions

Phase 3 – Open-Ended (Model + Web Search); 399 Question Subset



Open-ended evaluation reveals that while LLMs have a better chance of answering early-career level questions, their performance weakens on expert-level ones.

<u>This gap underscores the need for SME oversight and careful validation in critical utility applications.</u>

*__Difficulty-weighted__ scores reported to tighten dispersion across easy/medium/hard questions using weights of 1, 2, and 3 for easy/early-career, medium/experienced engineer, and hard/SME questions respectively. Each model was evaluated **three times** to measure run-to-run variability and ensure the results are statistically robust – **the median of the three is reported.**

Additional LLM Details: **Grok 4:** grok/grok-4-0709; **GPT 5**: openai/gpt-5-2025-08-07; **Gemini 2.5 Pro**: google/gemini-2.5-pro; **Claude Sonnet 4.5**: anthropic/claude-sonnet-4-5-20250929; **Perplexity Sonar Pro:** perplexity/sonar-pro

Benchmarking Demonstrates the Path to Trusted, Utility-Ready AI

Stakeholder Value

Utilities: Confidence in AI grounded in real-world, utility-specific benchmarks.
Vendors/Developers: Neutral evaluations that surface strengths and improvement areas.
Regulators & Policymakers: Independent standards to support safe AI adoption in critical infrastructure.

What's Next

Track Model Evolution: Continue benchmarking as LLMs improve, including domain-specific tools (e.g., EPRI.AI).
Shift to Use-Case Testing: Expand beyond generic tests into real utility applications (e.g., outage response, predictive maintenance, wildfire mitigation).

Image created using ChatGPT

Closing Thoughts

This work establishes the first domain-specific LLM benchmark for the electric power sector, advancing beyond MCQs to assess performance on real utility topics.

EPRI's benchmarking lays the foundation to evaluate domain-specific augmentation tools and models that can deliver greater value across the energy ecosystem.

www.epri.com