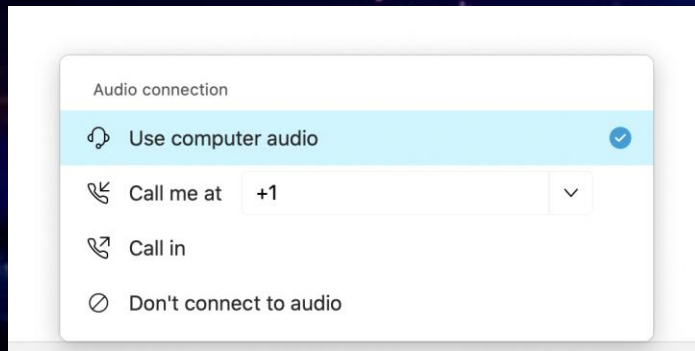
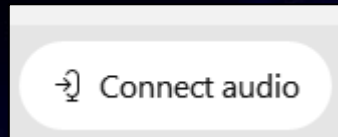


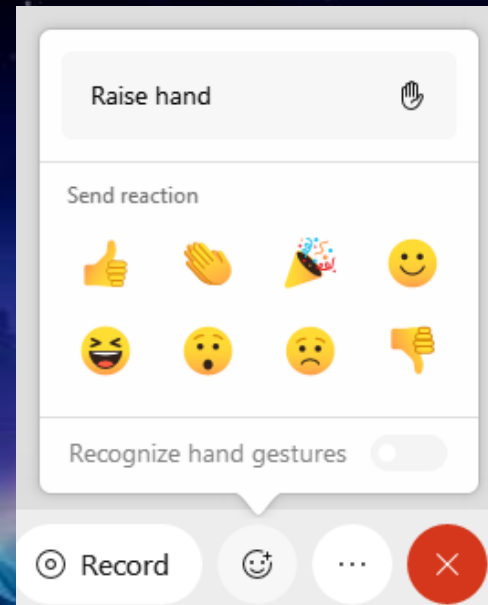
# Welcome to the Open Power AI Consortium Domain Specific Model Working Group

We will begin at 12:00 PM E.T.

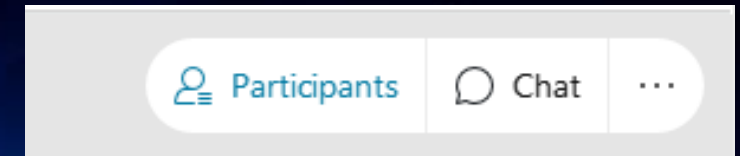
Be sure you have an audio icon next to your name. If you do not, disconnect your audio and select one of the options shown. Your line has been muted by the host.



Like what you see or hear or want to raise your hand? Let the Presenters know.



You can submit a written question at any time through the presentation using the CHAT panel. Please leave the recipient drop down to "Everyone".



In Chat:  
Name and Company

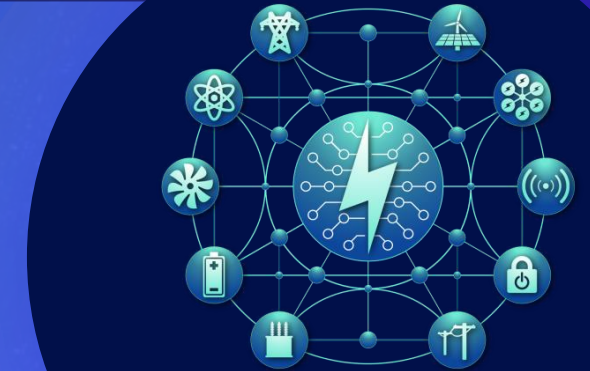


Reminder: This webcast is being recorded



# Open Power AI Consortium – Domain Specific Model Working Group

December Virtual Workshop



**OPEN POWER  
AI CONSORTIUM**

Ben Sooter

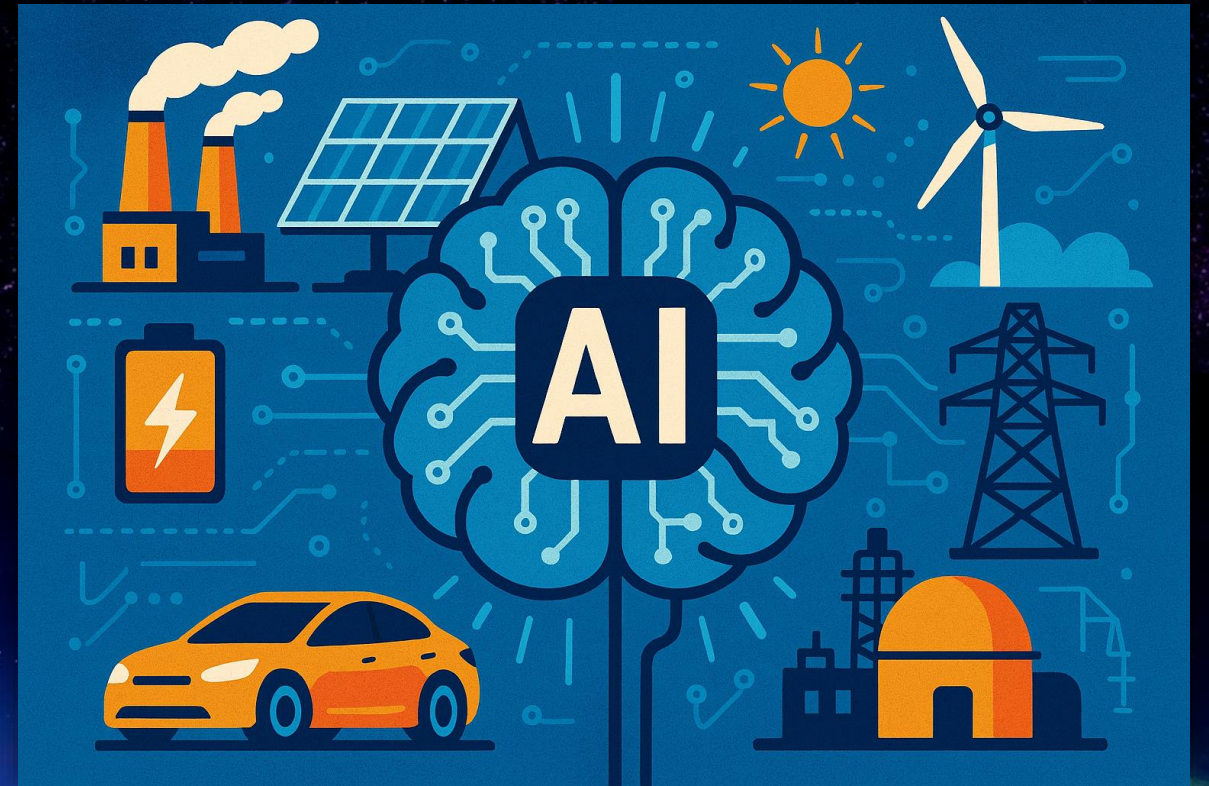
Program Manager

[bsooter@epri.com](mailto:bsooter@epri.com)

(p) +1-865-218-8108

# AGENDA

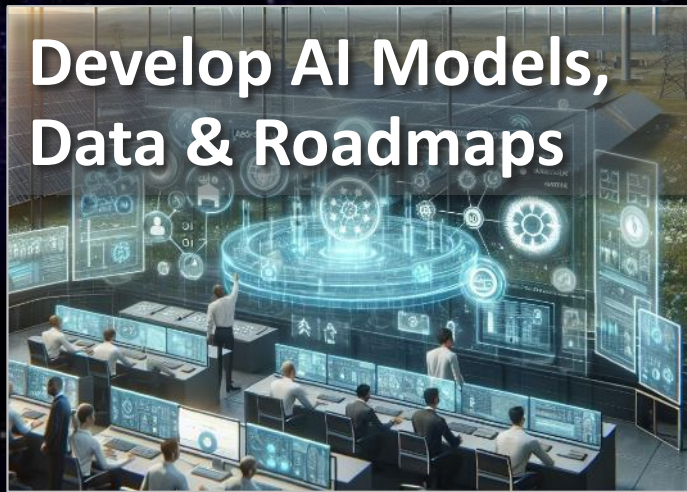
- Welcome and Safety Message
- Work Group Recap
- Upcoming Events
- Benchmarking Frontier LLMs
- DGX Spark
  - Running Models Locally On Your Desk
- RAG & Vector Stores



# SAFETY MOMENT



# Open Power AI Consortium – Key Objectives



## Develop AI Models, Data & Roadmaps

- Collaborate with leading AI developers to create AI models.
- Develop a feedback loop from deployments for refinement and optimization.
- Maximize impact via knowledge sharing among consortium members.

# DSM Workstream Goals & Desired Outcomes

## ▪ Workstream Purpose

- Develop and validate **Language Models purpose-built for the electric power sector**, capturing the terminology, data, and safety requirements unique to grid operations, planning, and compliance.

## ▪ Goals

- **Advance Domain-Specific LLMs** that outperform general models on utility tasks:
  - Dispatch, forecasting, compliance, and operator assistance, etc
- **Integrate AI safely into operational environments**, aligning with NERC CIP obligations.
- **Leverage Open Power AI infrastructure** to accelerate evaluation and deployment.
- **Collaborate across utilities, vendors, and labs** to build shared datasets, benchmarks, and evaluation frameworks.
- **Demonstrate measurable value** in reliability, operator efficiency, and compliance automation.

# Upcoming Events

OCTOBER						
Su	Mo	Tu	We	Th	Fr	Sa
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	







Work Stream Kickoff

NOVEMBER						
Su	Mo	Tu	We	Th	Fr	Sa
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

DECEMBER						
Su	Mo	Tu	We	Th	Fr	Sa
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

Benchmarking and Vector Store Development

Call for volunteers to present  
In January Meeting

	Implementation Work Stream Virtual Meeting
	Use-Case Work Stream Virtual Meeting
	Domain-Specific Model Virtual Meeting
	OPAI Executive Advisory Group
	OPAI Member Representative Committee
	EPRI Holiday

# Benchmarking Frontier Large Language Models (LLMs) for the Electric Power Sector

# Benchmarking Frontier Large Language Models (LLMs) for the Electric Power Sector



Apurba Sakti  
Principal Technical leader  
[asakti@epri.com](mailto:asakti@epri.com)  
(p) +1-704-595-2203

# Table of Contents



Image created using ChatGPT

This slide deck provides detailed results of EPRI's benchmarking effort. **For a high-level summary, view the white paper.**

## Motivation

- › Why Benchmark LLMs?

## Methodology & Implementation

- › EPRI's Benchmarking phases covered in this report
- › Generation of the Q&A datasets covering 35 power sector topics
- › Automated Evaluation using Inspect AI

## Results, Key Takeaways, & Next Steps

# 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.

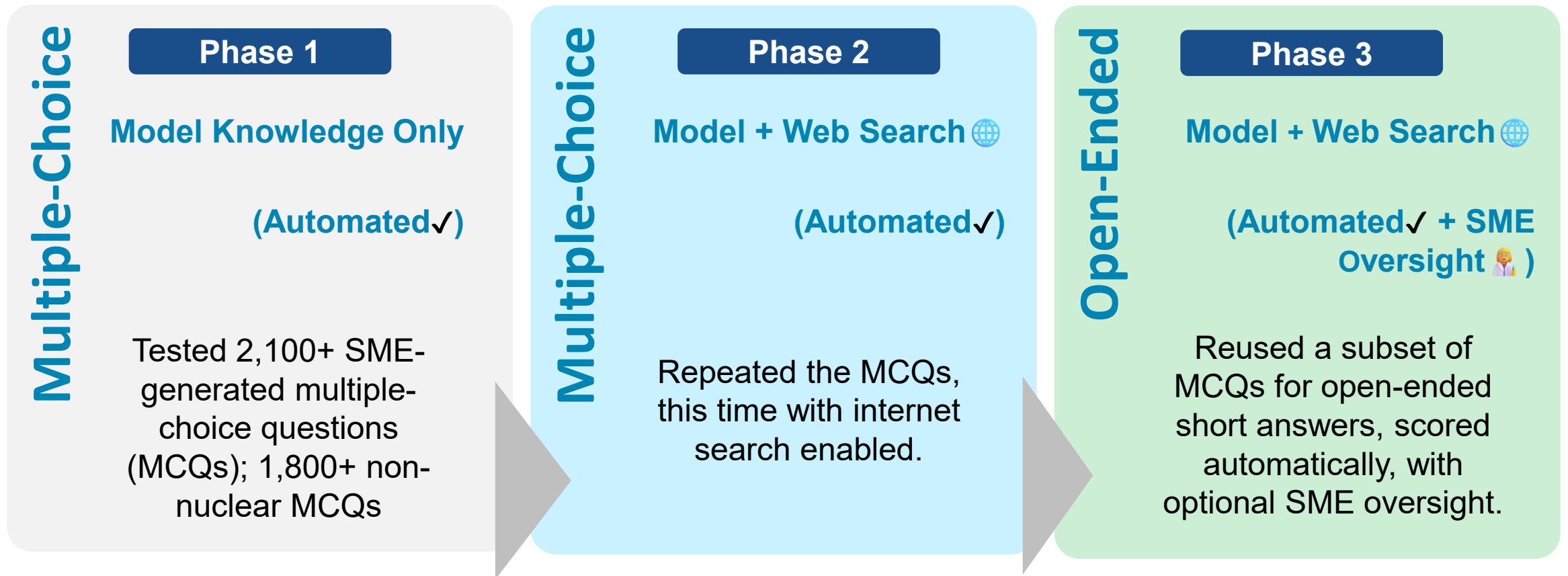
This gap motivates EPRI's multi-phase approach that evaluates large language models (LLMs) on over 35 power-sector topics. **This is the first step towards EPRI's evaluation of domain-augmented tools & real-world applications.**



# **Methodology & Implementation**

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

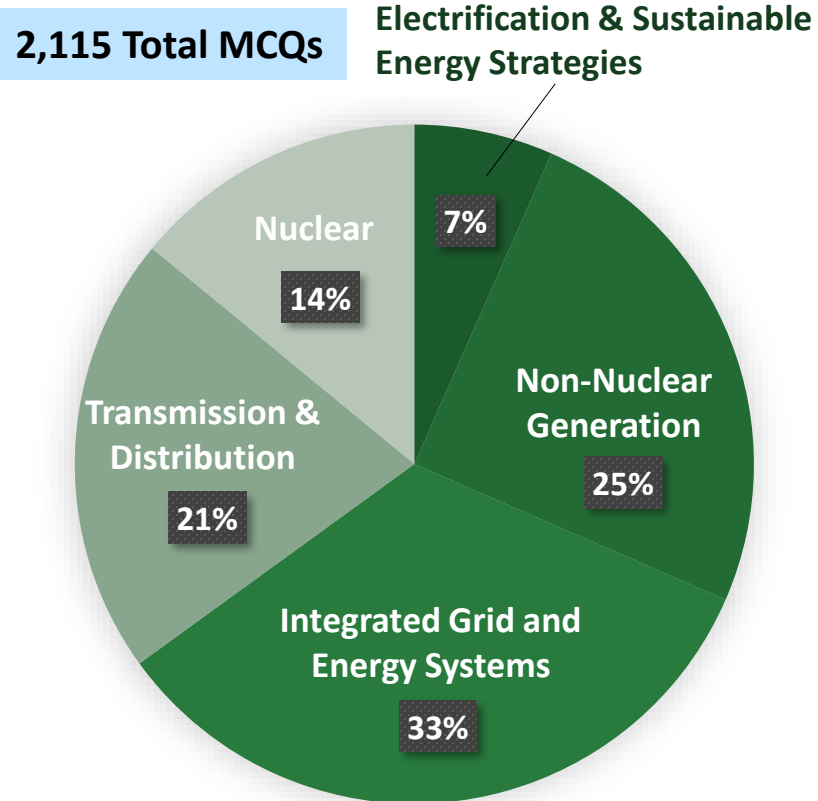
Phases 1-3 leverage automated repeatable evaluations and scoring with the option of SME oversight



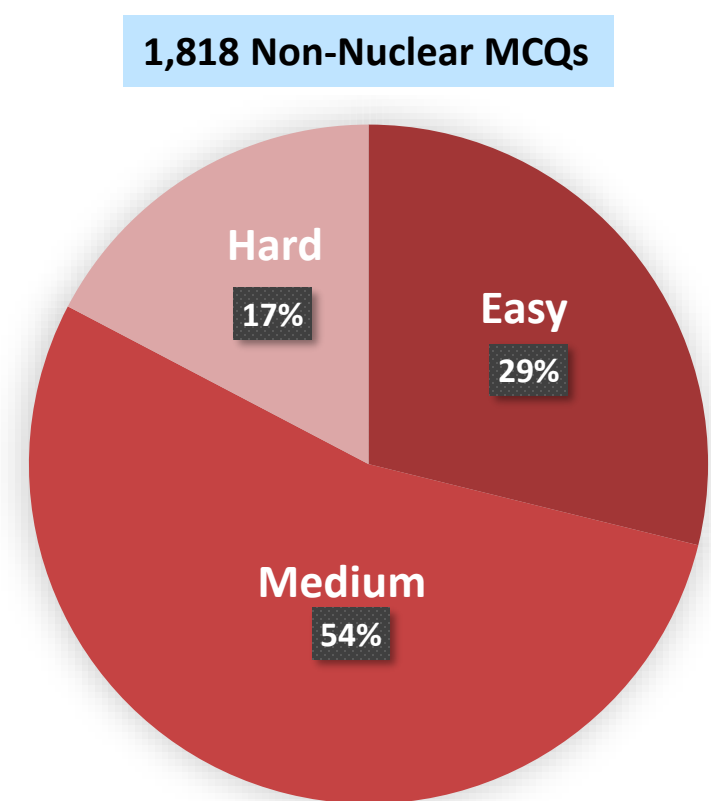
# EPRI's Subject Matter Experts developed a set of 2,115 Q&As

1,818 Q&As out of the 2,115 are on Non-Nuclear Topics

### Breakdown by Power Sectors



### Breakdown by Difficulty Level



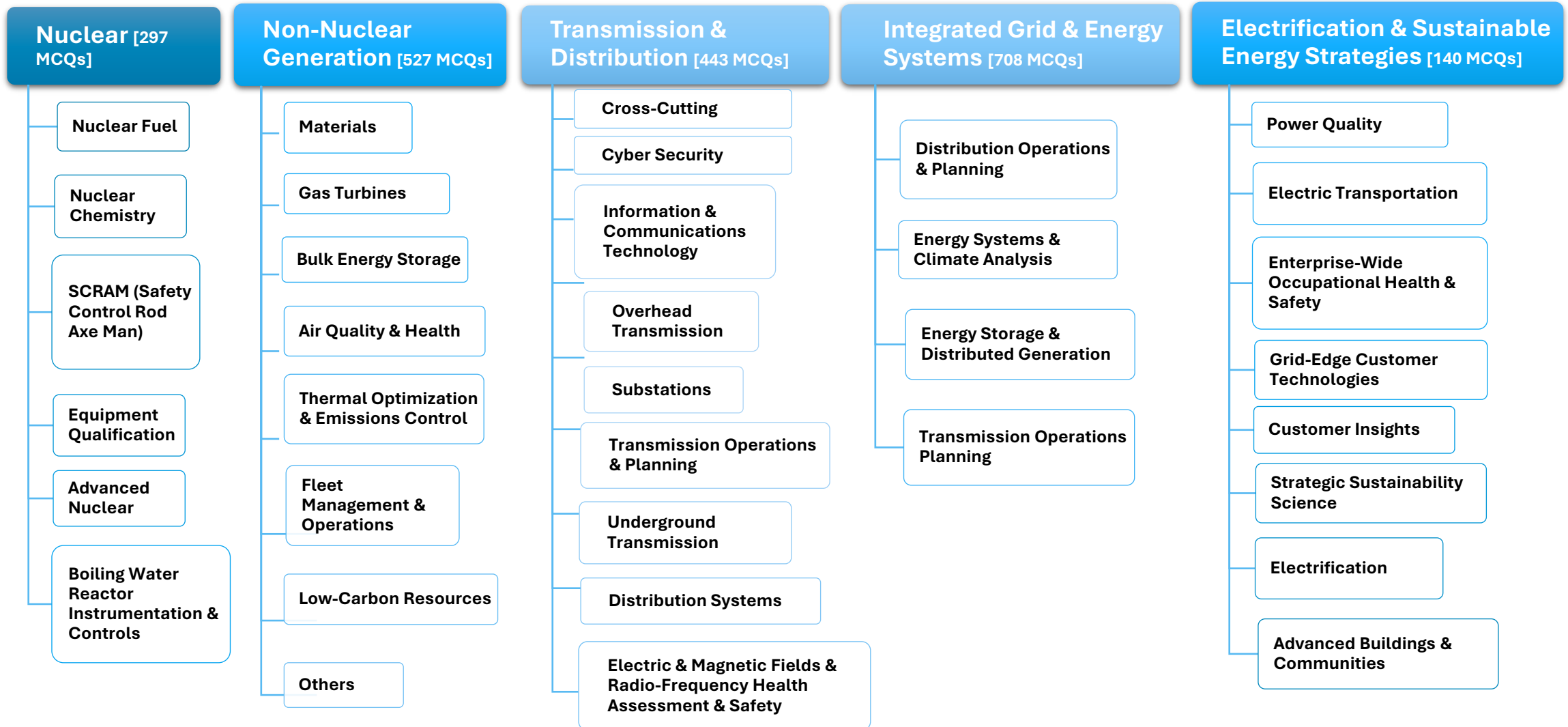
## Some Key SME Instructions

1. Prioritize publicly available data sources
2. Questions need to be formulated clearly and with context – the goal is not to trick the LLMs

\*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.

# EPRI's 2,115 Questions Span Major Power-Sector Domains

Multiple topics ensure coverage from nuclear & non-nuclear generation to grid operations & end-use applications



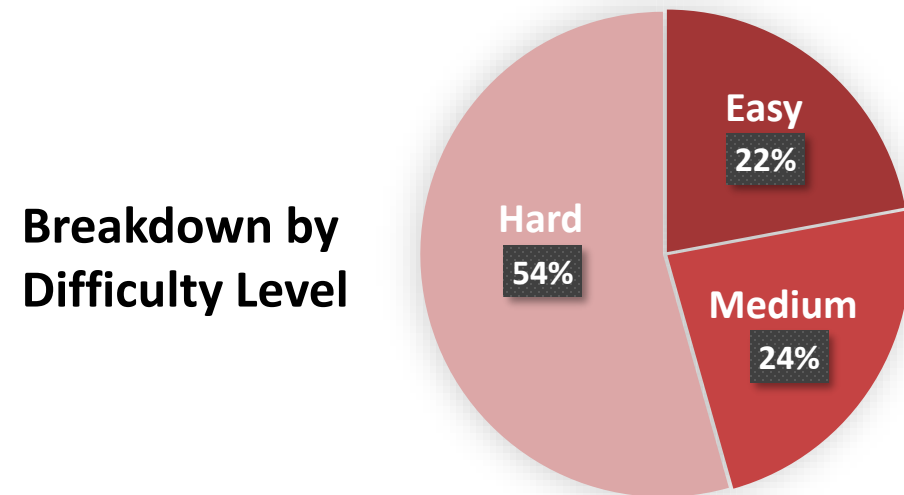
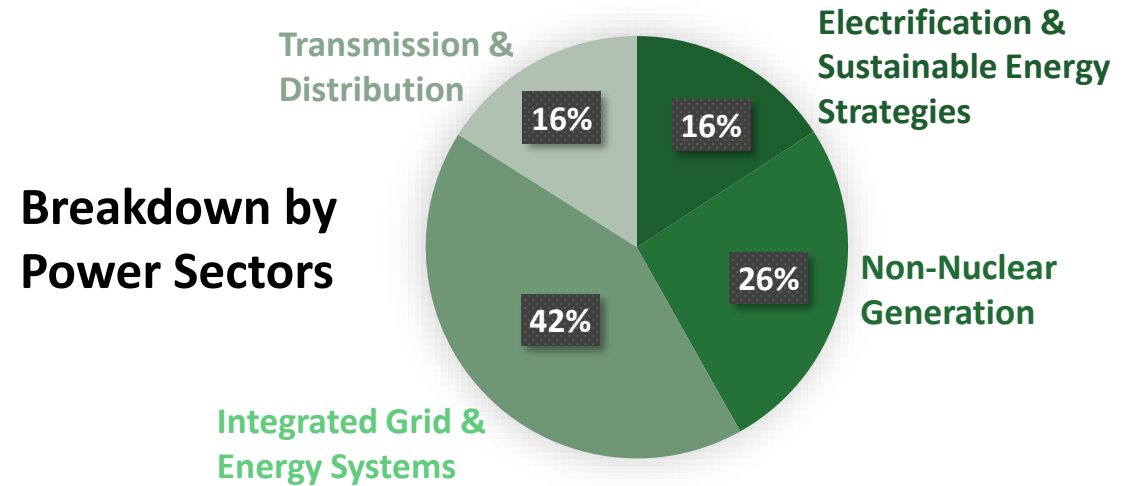
Due to the sensitive nature of nuclear content and restrictions on testing LLMs with nuclear-related material, this report primarily focuses on the 1,800+ non-nuclear Q&As.

# A subset of 399 non-nuclear questions were identified for open-ended evaluation in Phase 3

## 399 Questions for Open-Ended Evaluation

Not all MCQs can be converted to open-ended question format (e.g., “All of the above” questions)

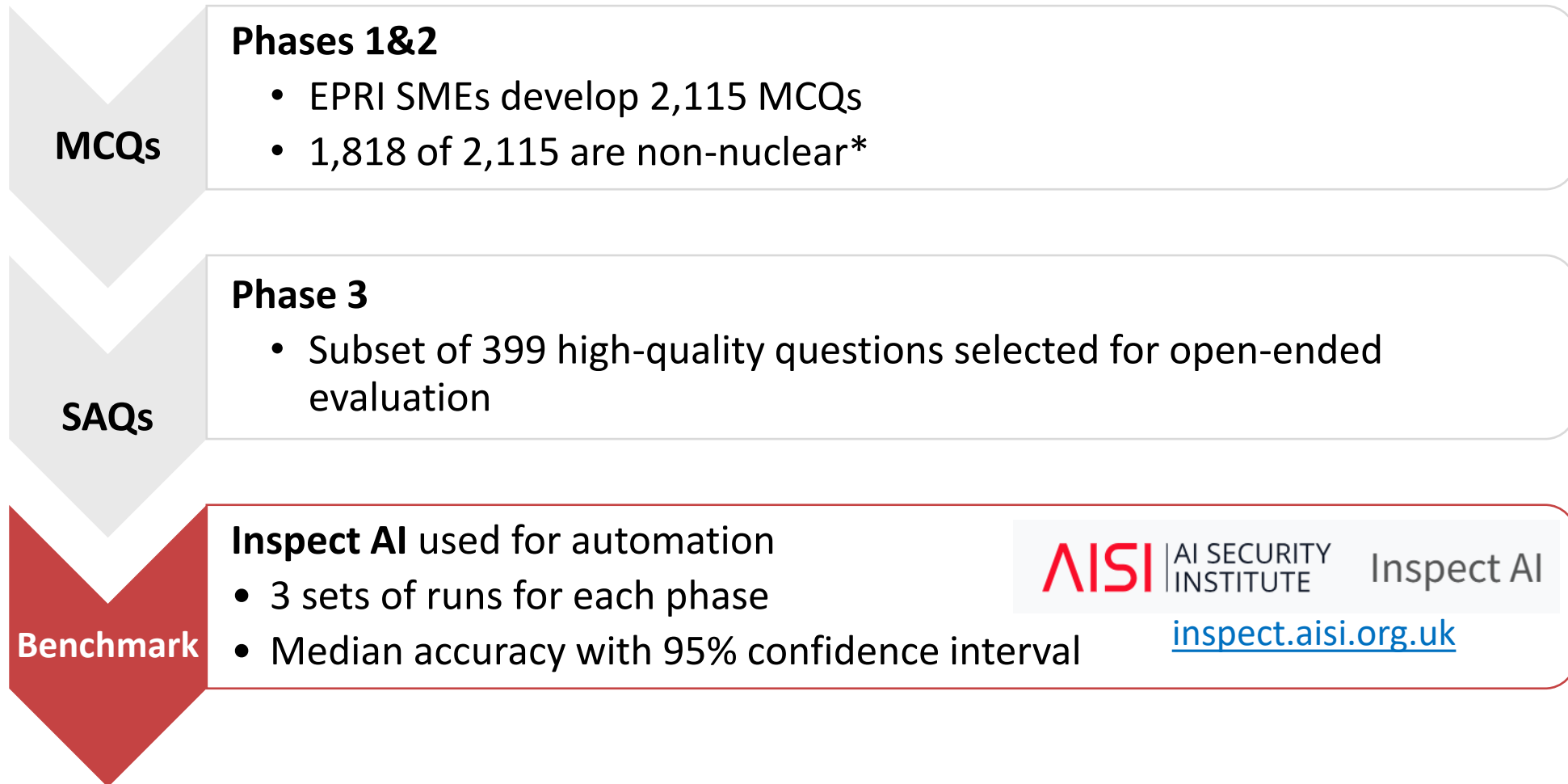
EPRI self-hosted LLMs were used to assess how well the 1,818 MCQs would translate to open-ended questions.



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

# 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: 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.



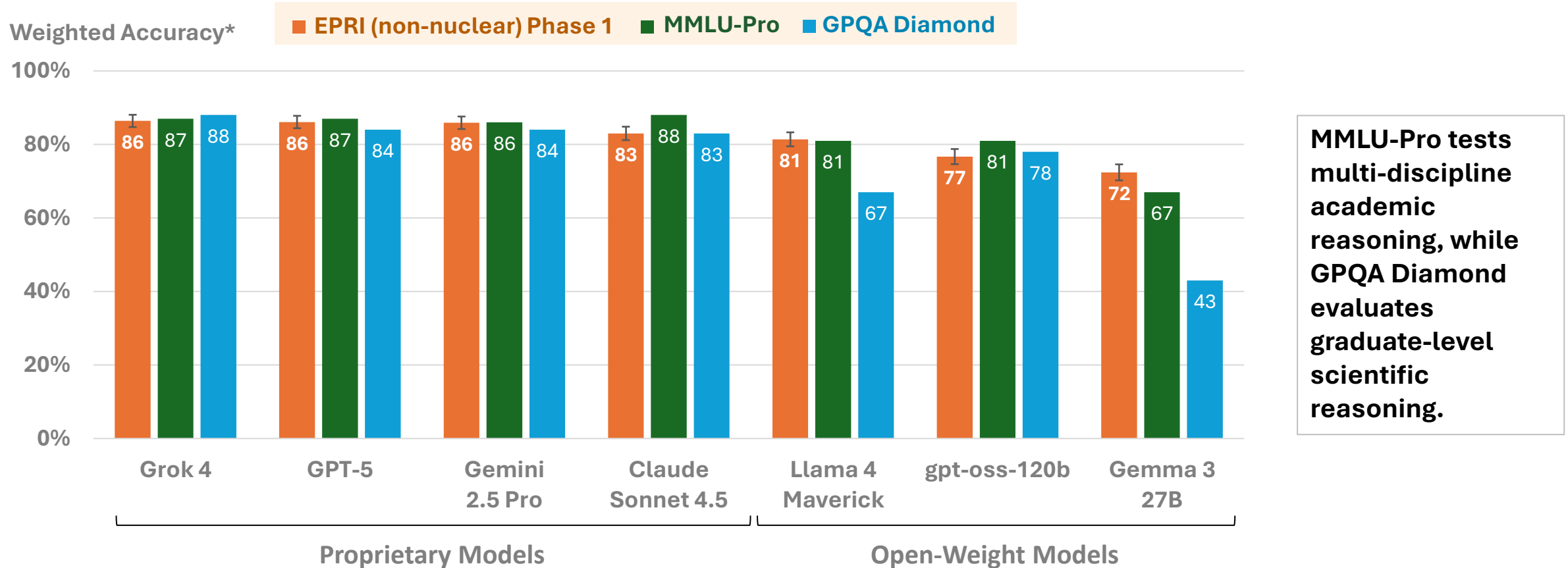
# **Results, Key Takeaways, & Next Steps**

# Model Accuracies Generally in Alignment with External Benchmarks

Grok 4, GPT 5, and Gemini 2.5 Pro among the top AI large language models across all 3 benchmarks

## EPRI's Power-Systems Benchmarking vs. External Graduate-Level Benchmarks

1,818 Multiple Choice Questions (Model Knowledge); 95% Confidence Intervals



\***Difficulty-weighted** scores reported for EPRI evaluations 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](#).** MMLU Pro, and GPQA Diamond scores from [Artificial Analysis](#).

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; **Llama 4 Maverick:** azureai/Llama-4-Maverick-17B-128E-Instruct-FP8; **got-oss-120b:** epri-hosted/openai/gpt-oss-120b; **Gemma 3 27B:** epri\_hosted/google/gemma-3-27b-it

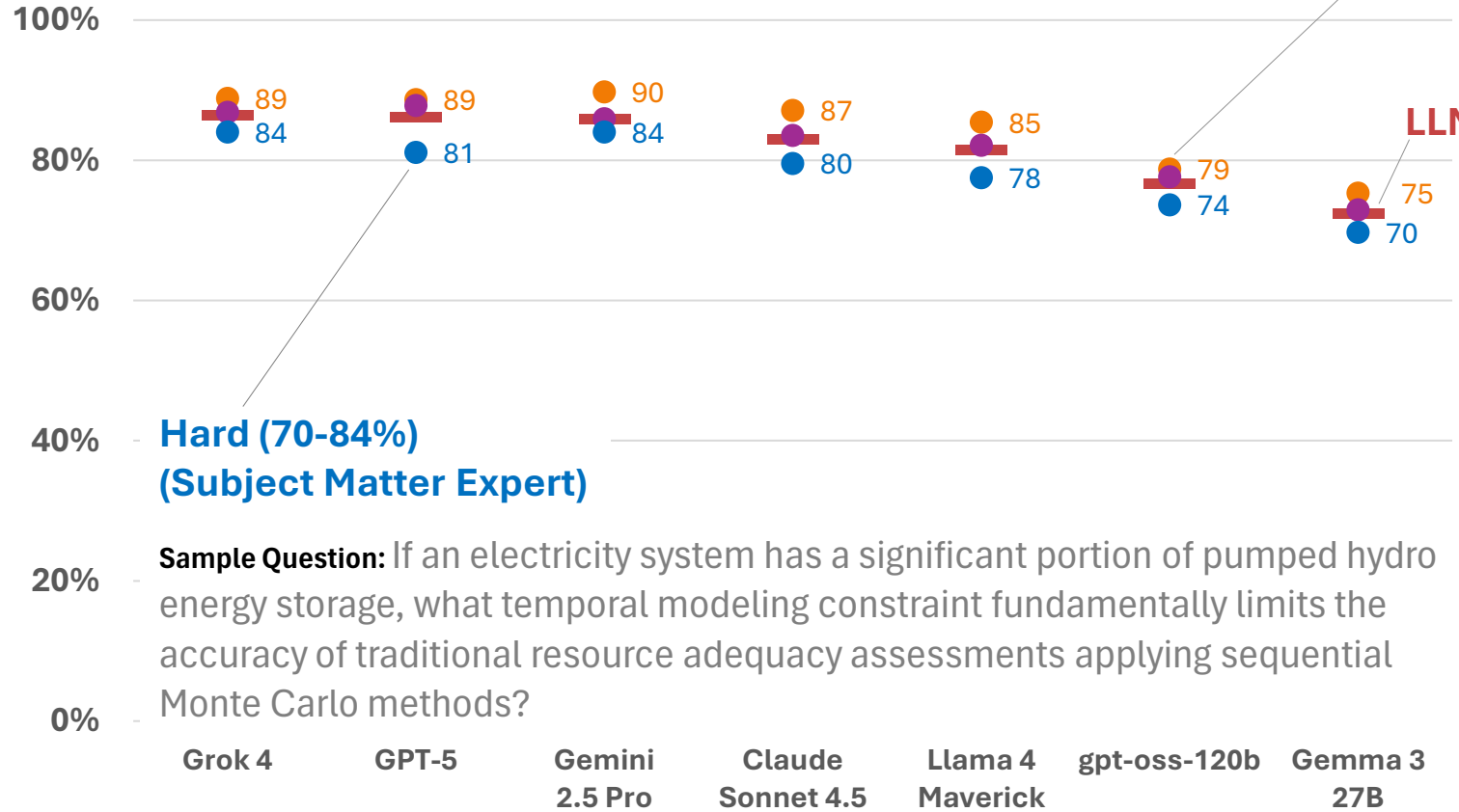
# Frontier Models Scored Lower (70-84%) on Hard MCQs

For easy questions, these models scored between 75-90%

## Model Performance Across Easy, Medium, & Hard Questions

Phase 1 – Multiple Choice Questions (Model Knowledge); 1,818 Questions

Weighted Accuracy\*



Easy (75-90%)  
(Early Career)

Sample Question: For U.S. residential customers, what is the most common default type of electricity rate structure?

LLM Weighted Average Performance (52-67%)

Medium (73-88%)  
(Experienced Utility Engineer)

Sample Question: What limits multivariable regression analysis from establishing a correlation between distributed generation (e.g., solar) and customer adoption of electrification for future extrapolation?

Hard (70-84%)  
(Subject Matter Expert)

Sample Question: If an electricity system has a significant portion of pumped hydro energy storage, what temporal modeling constraint fundamentally limits the accuracy of traditional resource adequacy assessments applying sequential Monte Carlo methods?

\*Difficulty-weighted scores reported for EPRI evaluations 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; **Llama 4 Maverick:** azureai/Llama-4-Maverick-17B-128E-Instruct-FP8; **got-oss-120b:** epri-hosted/openai/gpt-oss-120b; **Gemma 3 27B:** epri\_hosted/google/gemma-3-27b-it

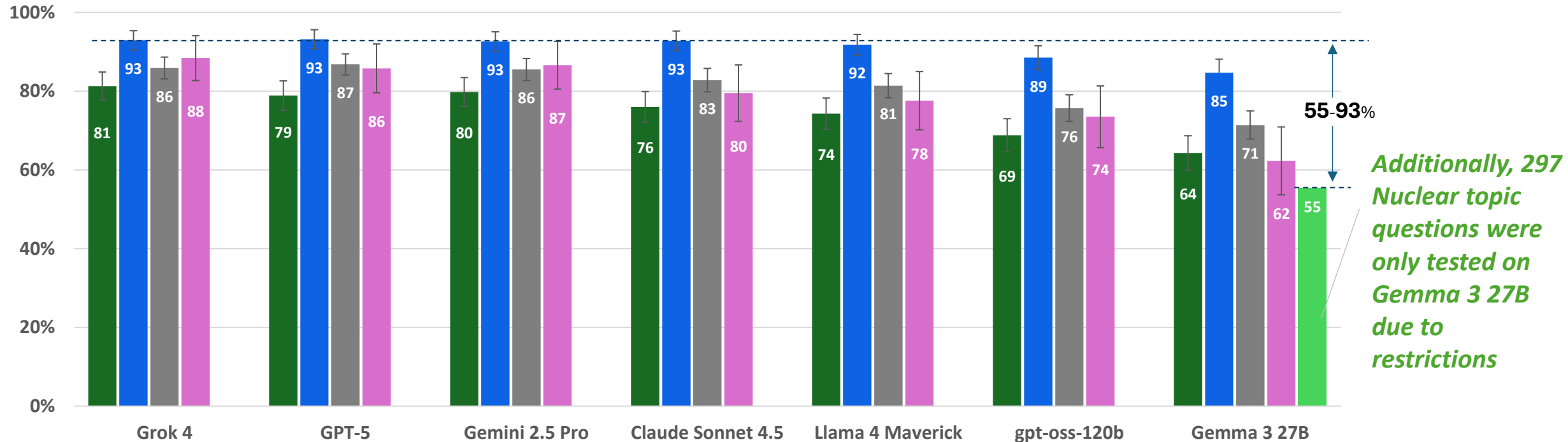
# Accuracy Scores Varied from 64-90% Across Non-Nuclear Topics

EPRI-hosted Gemma scored 55% on Nuclear questions

**LLM Model Accuracy Across a Range of Power System Topics**  
**1,818 Multiple Choice Questions (Model Knowledge); 95% Confidence Intervals**

■ Topics on **non-nuclear generation** ■ Topics on **transmission and distribution** ■ Topics on **integrated grid & energy systems**  
 ■ Topics on **electrification & sustainable energy strategies** ■ Topics on **nuclear energy**

Weighted Accuracy\*



\***Difficulty-weighted** scores reported for EPRI evaluations 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](#).**

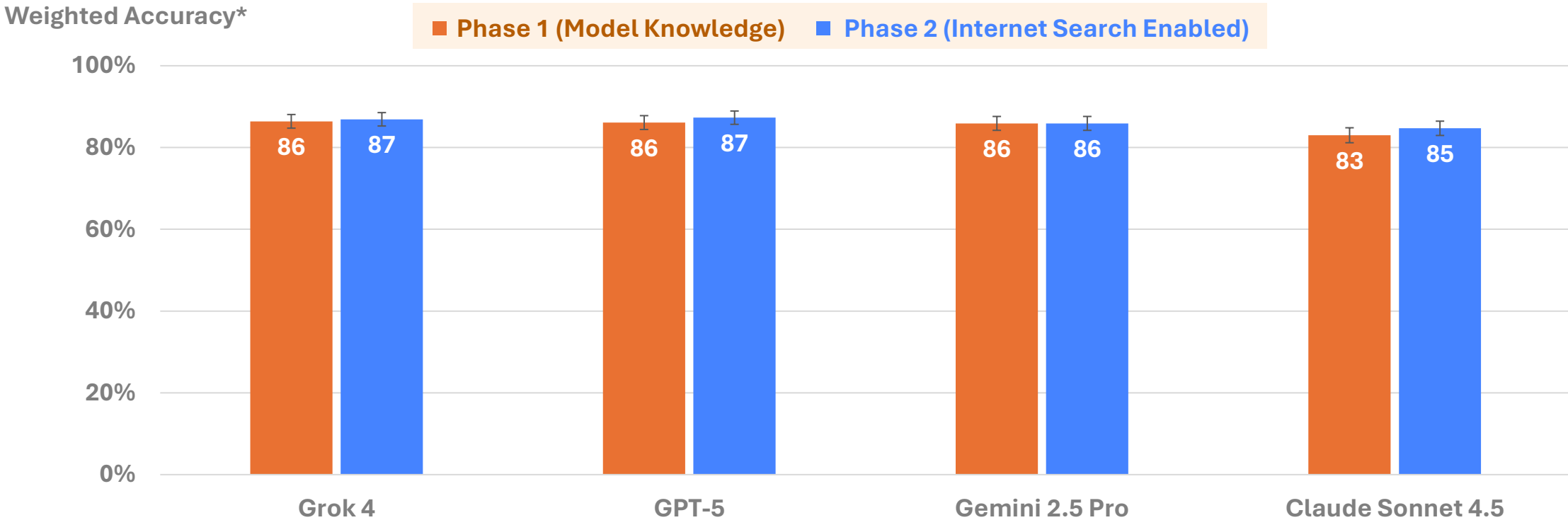
Additional LLM Details: **Grok 4:** grok/grok-4-0709; **GPT 5:** openai/gpt-5-2025-08-07; **Gemini 2.5 Pro:** google/gemini-2.5-pro; **Claude Sonnet 4.5:** anthropic/claude-sonnet-4-5-20250929; **Llama 4 Maverick:** azureai/Llama-4-Maverick-17B-128E-Instruct-FP8; **got-oss-120b:** epri-hosted/openai/gpt-oss-120b; **Gemma 3 27B:** epri\_hosted/google/gemma-3-27b-it

# For Multiple-Choice Questions, Enabling Internet Search Has a Minimal Effect on Accuracy

Scores seen to improve by up to 2pts.

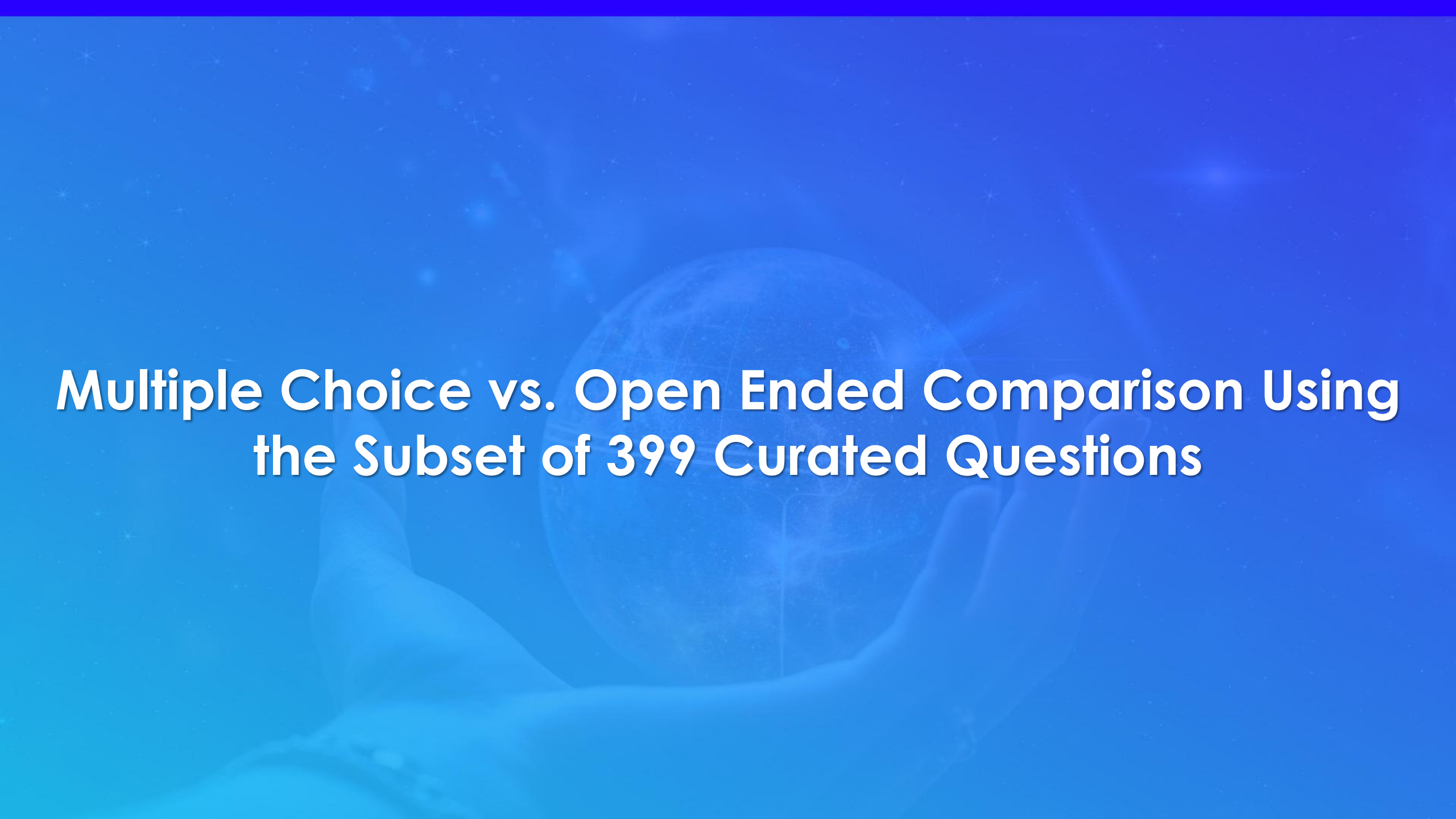
## EPRI's Power System Benchmarking

1,818 multiple choice questions (MCQs) with and without internet search; 95% Confidence Intervals



\*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



**Multiple Choice vs. Open Ended Comparison Using  
the Subset of 399 Curated Questions**

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

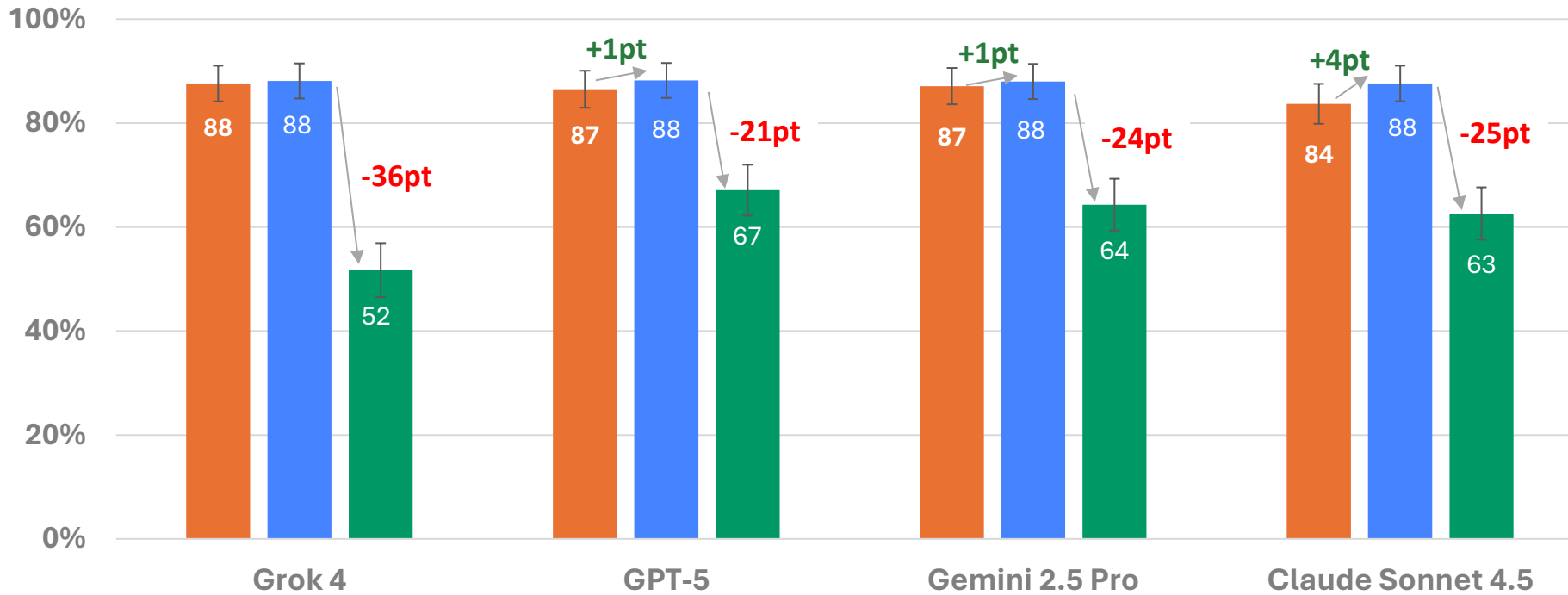
Some frontier LLMs answered incorrectly ~50% of the time

## 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)

Weighted Accuracy\*



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](#)

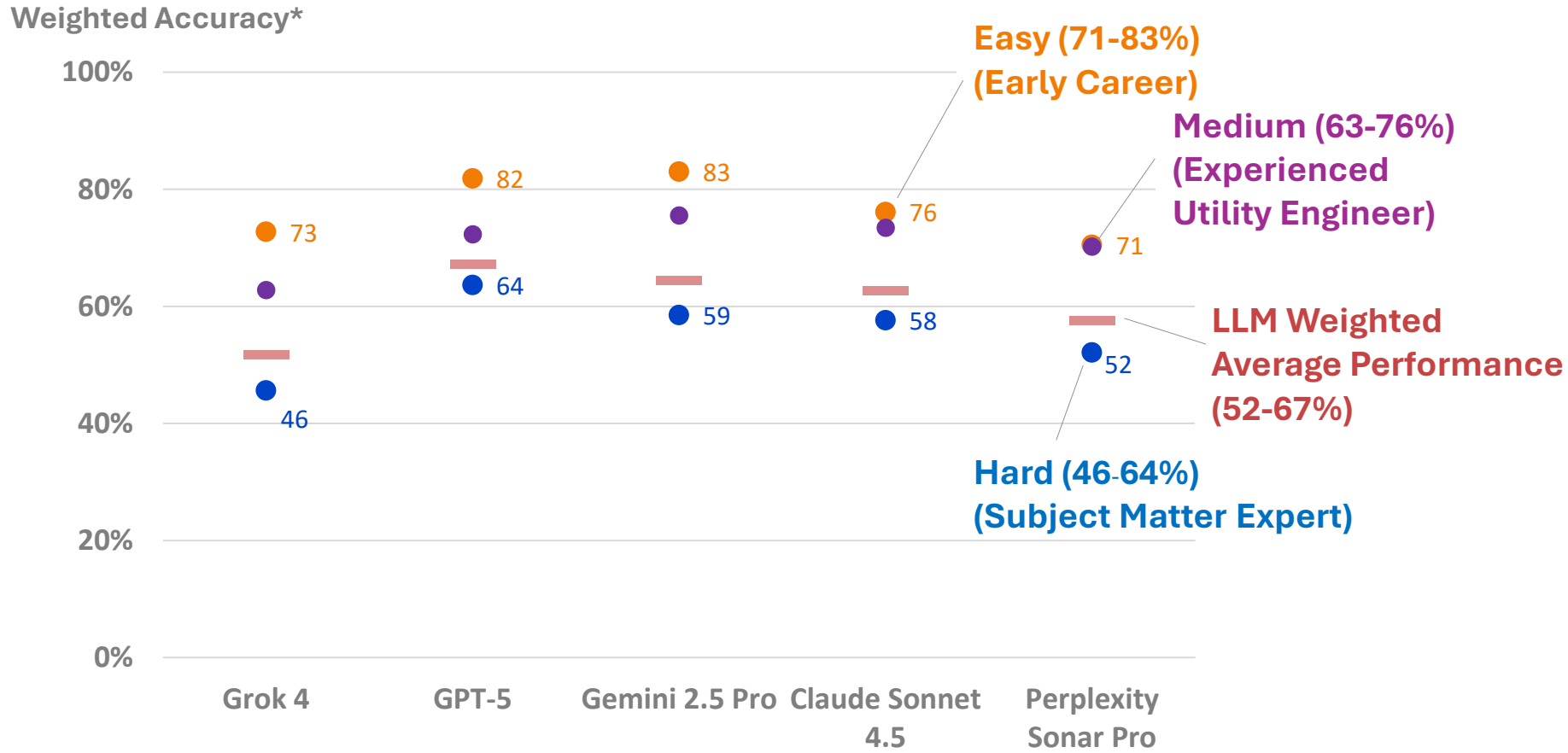
Additional LLM Details: **Grok 4:** grok/grok-4-0709; **GPT 5:** openai/gpt-5-2025-08-07; **Gemini 2.5 Pro:** google/gemini-2.5-pro; **Claude Sonnet 4.5:** anthropic/claude-sonnet-4-5-20250929

# 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.

**This gap underscores the need for SME oversight and careful validation in critical utility applications.**

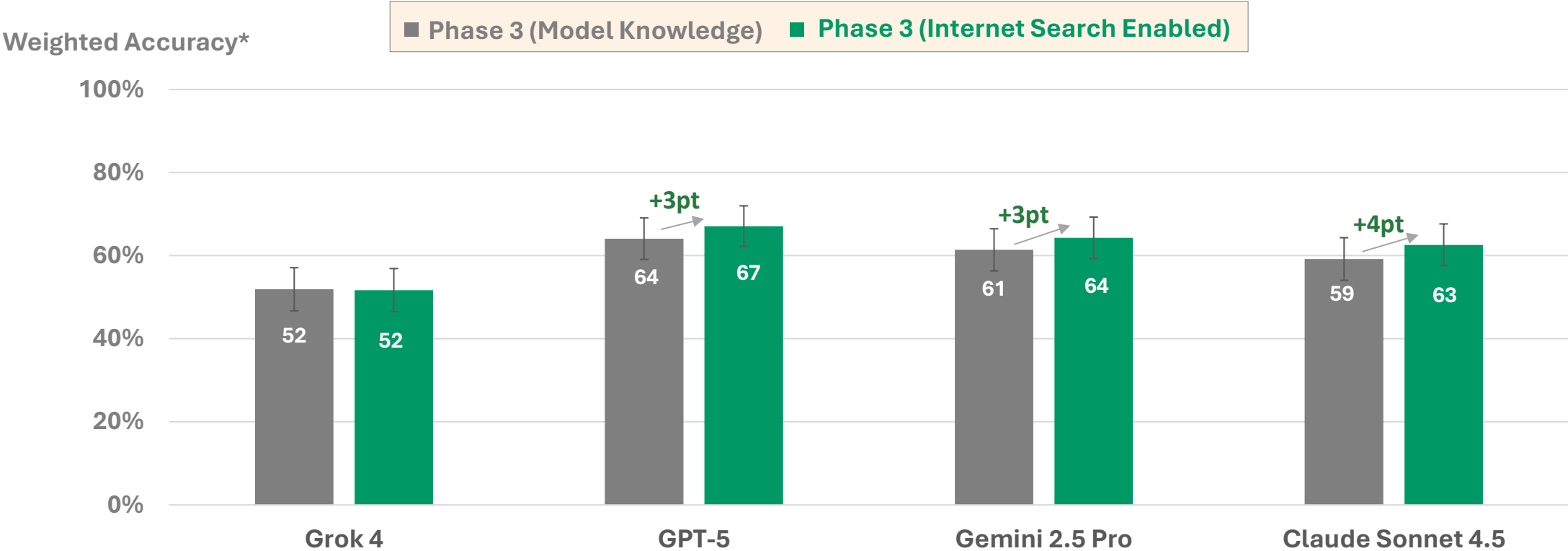
\***Difficulty-weighted** scores reported to tighten dispersion across easy/medium/hard questions using weights of 1, 2, and 3 for easy/early-career, medium/experienced engineer, and hard/SME questions respectively. Each model was evaluated **three times** to measure run-to-run variability and ensure 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

# For Open-Ended Questions, Internet Search Improved LLM Accuracy by up to 4pts. vs up to 2pts. for MCQs

## EPRI's Power System Benchmarking

Phase 3 open-ended short-answer questions (SAQs) with and without internet search, with 95% Confidence Intervals



\*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

# 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.**

# DGX Spark

# DGX Spark in 100 seconds

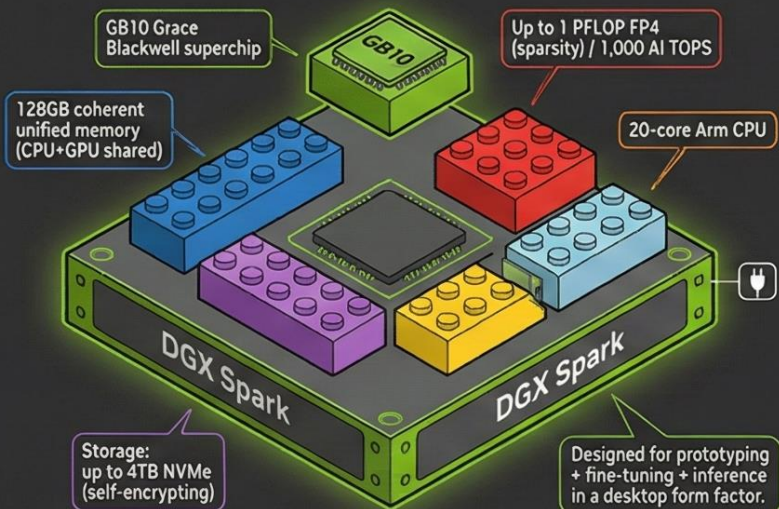


Image source: NVIDIA

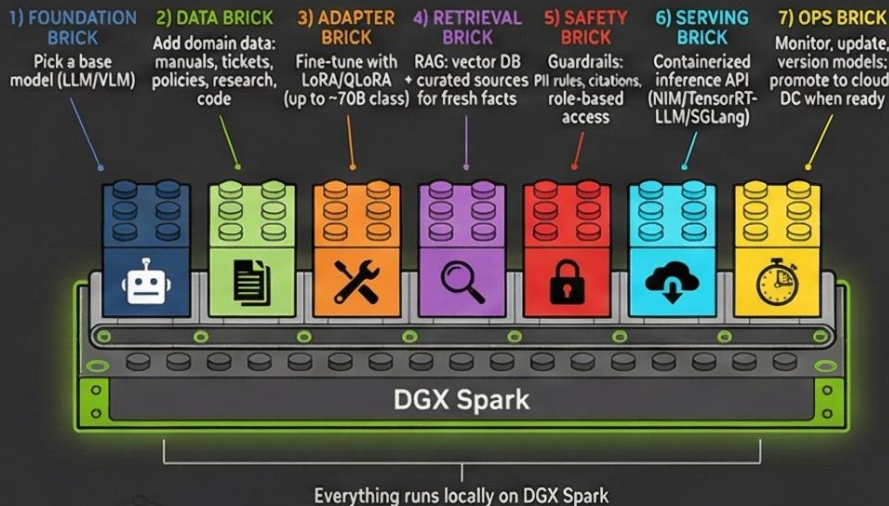
# NVIDIA DGX Spark: Snap-Together Domain AI on Your Desk

Build, fine-tune, and serve domain-specific models locally—then ship the same stack to the data center/cloud when ready.

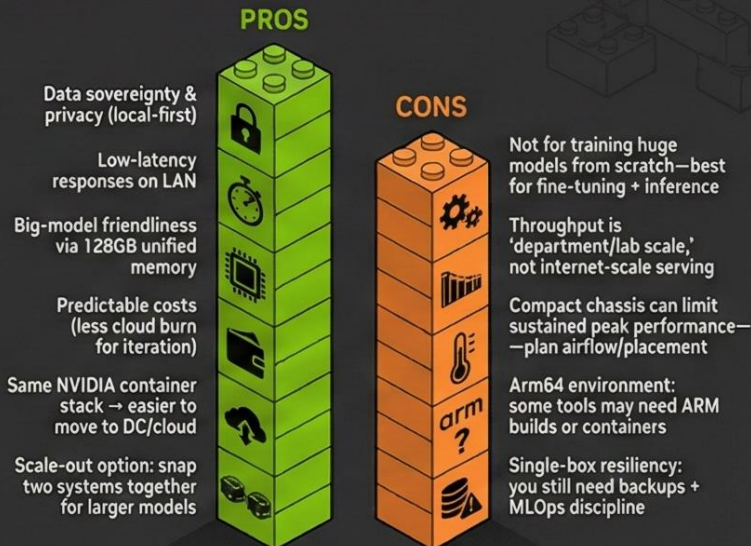
## 1 WHAT IS DGX SPARK?



## 2 HOW DOMAIN-SPECIFIC MODELS GET BUILT (BRICK BY BRICK)



## 4 PROS vs CONS



## 3 REALISTIC SMALL-DESKTOP USE CASE BUILDS



**A) Private Document Copilot** RAG over internal docs + domain LoRA; keeps sensitive data on-prem.

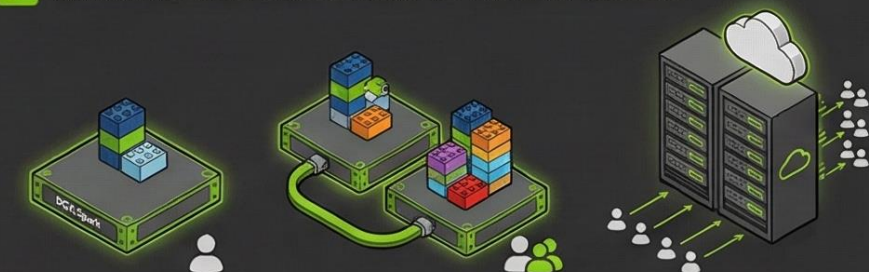
**B) Customer Support Copilot** Summarize tickets, draft replies, suggest macros; integrates with CRM via tools.

**C) Manufacturing / Field-Service Assistant** Troubleshooting from manuals + maintenance logs; generates step-by-step checklists.

**D) Security Operations Helper** Summarize SIEM alerts, map to runbooks, draft incident notes (local-first for sensitive logs).

**E) Edge Vision + Robotics Prototyping** Prototype Metropolis/Isaac/Holoscan pipelines; iterate locally before deployment.

## 5 WHEN TO USE ONE SPARK vs TWO vs 'GO BIG'



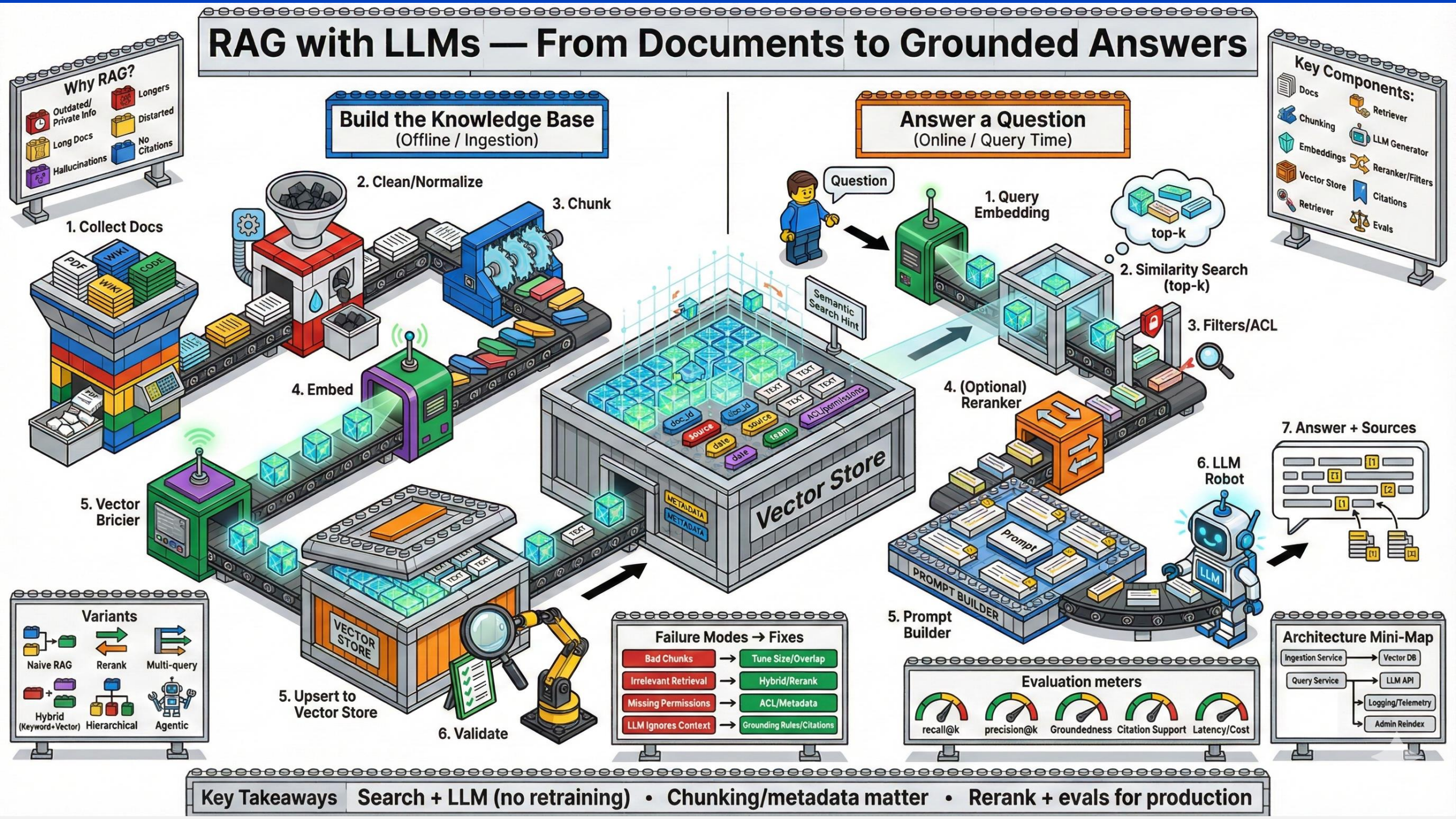
**ONE SPARK**  
Single team, sensitive data, rapid prototyping, local inference

**TWO SPARKS**  
Bigger models & parallel experiments

**DATA CENTER/CLOUD**  
High-throughput inference, HA requirements, large-scale training

# Retrieval-Augmented Generation

# RAG with LLMs — From Documents to Grounded Answers



# Discussion Questions

# TOGETHER...SHAPING THE FUTURE OF ENERGY®



**OPEN POWER  
AI CONSORTIUM**