

Housekeeping

A few norms to keep today's discussion useful, interactive, and publicly shareable

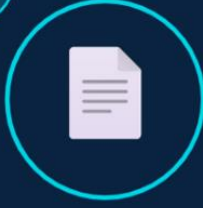
1



Participate actively

Please mute upon entry. Use raise hand or chat for questions and discussion.

2



Keep it publicly shareable

Please share only material suitable for public release.

3



No automated notetakers

Please do not add automated notetakers to this meeting.

4



Recording notice

This session will be recorded. Participation acknowledges consent.

5



Goal for today

Practical discussion and shared learning. Product pitches should be avoided or kept at a minimum.

Building Domain Intelligence

What We're Learning, What Comes Next

A working session on the choices that shape domain-model quality.



Apurba Sakti, *Ph.D.*

Electric Power Research Institute

Domain Intelligence Working Group | August 20, 2026

PASSWORD ADVICE HAS EVOLVED

SAFETY SHARE MOMENT

NIST GUIDANCE

Don't require routine resets.

Change only if compromised.



WHY IT MATTERS

97%

of identity attacks observed by Microsoft in 2025 were **password-spray attacks**.

BETTER AUTHENTICATION

Passkeys are phishing-resistant and tied to the real website



WHAT TO DO



Use long, unique passwords.



Use a passkey when available.



Be alert for sign-in prompts you didn't start.

Sources: NIST SP 800-63B-4, Passwords section: <https://pages.nist.gov/800-63-4/sp800-63b.html>; Microsoft Digital Defense Report 2025: <https://www.microsoft.com/en-us/corporate-responsibility/topics/cybersecurity/reports/microsoft-digital-defense-report-2025/>; Fast Identity Online: How Passkeys Work <https://fidoalliance.org/how-fido-works/>



Powered by EPRI

OPEN POWER AI CONSORTIUM

Our Work Groups

Domain Intelligence



Enabling domain-aware AI performance for power sector applications

Use Case Sandbox



Identify, prioritize, and evaluate real-world applications

Implementation



Share methods & best practices for AI deployment and scaling



Harnessing AI's transformative potential for a resilient, cost-effective energy future through collaboration

Open-source data and AI models for industry-wide value



Google



ORACLE



Learn More:
www.openpowerai.org



Domain-Specific Model
to include knowledge from

~ 14,000

EPRI Reports



250+

use cases identified for consortium prioritization and development

Join us in Shaping the Future of AI and Energy

Today's throughline: from identifying gaps to building domain intelligence

Disciplined choices can connect to more reliable power-sector AI



Framing the session

- Benchmarking identifies where domain capability is still needed
- It further allows us to track whether the gaps are being filled as domain adaptation is undertaken



Shammya Saha, Ph.D.
Staff Engineer, EPRI

From source material to usable supervision

Preserve technical structure and meaning

Question design | Answer targets
| Reasoning traces



Chris Trueblood
Principal Technical Leader, EPRI

From domain exposure to demonstrated capability

Match the adaptation method to the objective

CPT / SFT / SDFT / RL | Train-eval | Retention

Working Group Discussion

- Design choices that have greatest downstream impact
- Lessons to share from the audience
- Topics to explore next

Takeaway

Domain intelligence depends not only on where we train, but also on what we teach, how we adapt, and how we verify the result.

Finding the gap

WattWorks.EPRI.Com
The Power Sector's AI Benchmarking Hub

Coming Next:
Agentic AI Evaluations

The Power Sector's AI Benchmarking Hub

How Today's Leading LLMs Perform on Power-Sector Questions

EPRI benchmarks leading large language models on **real-world, utility-relevant questions** to help the power sector understand where today's AI systems are strong, **where they fall short**, and how performance evolves over time.

● Live Benchmark — Updated as new models are evaluated

Download White Paper →

Watch Benchmarking Discussion →



Multiple-choice evaluations



Open-ended short answers



Multi-run repeatability



Domain-augmented tools

INTERACTIVE BENCHMARK

Explore how selected models perform across EPRI benchmark datasets, including **short-answer** and **multiple-choice questions**, **model-only** and **web-search modes**, and breakdowns by **difficulty** or **topic**.

DATASET ⓘ

☒ Multiple-Choice Questions (MCQ)

☐ Short-Answer Questions (SAQ)

MODE ⓘ

☒ Search Off (Model Only)

☐ Search On (Web Search)

VIEW ⓘ

☒ Overall

☐ By Difficulty

☐ By Topic

MODELS ⓘ 5/10 selected Select default 10 Clear all

☐ Grok 4

☒ GPT-5

☐ Gemini 2.5 Pro

☐ Claude Sonnet 4.5

☐ Llama 4 Maverick

MCQ — Overall

Weighted Accuracy (%)

Search Off (Model Only)

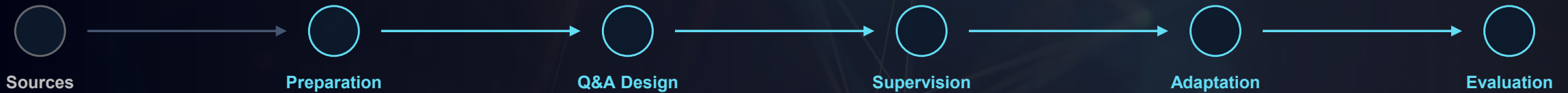


About these results

Scores reflect difficulty-weighted accuracy across EPRI's multiple-choice benchmark set. Higher scores indicate better performance. Error bars show 95% confidence intervals where available. Search Off uses model knowledge only, without web search.

Finding the gap is only the beginning

The next phase is deciding what to train on, how to teach it, and how to evaluate what changes.



Domain intelligence is shaped long before, and after, the training run.

Preparation

**PDF → Markdown
can alter structure**

Headings, equations,
tables, or reading order can
be lost.

Q&A Design

**Q&A quality depends
on careful design**

Ambiguity or missing context
leaves several defensible
answers.

Supervision

**Teacher traces need
validation**

Plausible traces may be
speculative; only grounded
reasoning should
supervise training.

Adaptation

**Match method to
objective**

CPT, SFT, SDFT, and
post-training change
different capabilities.

Evaluation

**Improvement can be
hard to interpret**

Overlap, memorization,
judge behavior, or format
effects can distort results.

Any examples to share on what has resulted in the greatest downstream impact for you?

Document preparation determines what structure reaches the model

Training-ready content should preserve the structure and context that carry technical meaning.



Raw extraction (structure flattened)

voltage impacts reverse flow pv

table 2.2 results feeder peak vmin vmax pass

a	4.8mw	0.93	1.07	no
b	3.1mw	0.95	1.04	yes

figure 4 below shows profile

page 17

Hierarchy, tables, and figure context are hard to recover from flat text.

Preparation makes structure usable

- 1 Section hierarchy restored -----
- 2 Table structure preserved -----
- 3 Figure linked to caption -----
- 4 Page/header noise removed -----



Training-ready representation (structure preserved)

Voltage impacts under reverse flow

Table 2.2 – Feeder voltage screening

Feeder	Peak PV	Vmin	Vmax	Result
A	4.8 MW	0.93	1.07	Review
B	3.1 MW	0.95	1.04	Pass

Figure 4: Voltage profile at peak PV output.



Good preparation keeps technical structure and context intact.

Good questions test domain reasoning, not interpretation

The difficulty should sit in the engineering, not in guessing what was meant.

Underspecified prompt

A PV inverter is running at unity power factor and the voltage at its connection point is a bit high. Should it start absorbing reactive power to fix this?

Why it fails

The model requires guessing the asker's intent, so the difficulty sits in the prompt rather than in the grid reasoning

What's missing

inverter limits • location • target criterion

Well-bounded prompt

A 500-kW PV inverter (pf range 0.8 lagging to 0.8 leading) at the end of a rural feeder runs at unity power factor. At noon the PCC voltage rises to 1.06 pu, above the ANSI C84.1 limit. A technician proposes changing the inverter's power-factor setpoint. Should the inverter absorb or inject reactive power to bring the voltage back in range, and why does operating at unity PF leave voltage support 'on the table'?

Expected target

The fix is to move off unity PF to a Q-absorbing setpoint. Unity PF forfeits this because $Q=0$ means no reactive lever at all

What's specified

equipment limits • location • operating condition • target criterion

Where is ambiguity worth keeping because it mirrors how the decision is really made?

Sources

Preparation

Q&A Design

Supervision

Adaptation

Evaluation

Designing the answer matters as much as designing the question

A complex reasoning can still have a clear, checkable target answer.

Task

Given abnormal feeder voltage after a large PV interconnection, identify the most likely first area to investigate and explain why.

Verbose target

“It depends on the feeder, but the utility could consider voltage regulators, inverter settings, capacitor bank changes, protection updates, hosting capacity limits, or other mitigation depending on local conditions...”

- Hard to score consistently
- Many plausible paths
- Verbosity can mask weak reasoning

Constrained target

“First check voltage-control settings and volt/VAR behavior against measured reverse-flow conditions; prioritize settings-based mitigation before new equipment if no thermal or protection constraint is present.”

- Clear conclusion
- Traceable rationale
- Cleaner supervision signal

A better answer is bounded, not longer

Sources

Preparation

Q&A Design

Supervision

Adaptation

Evaluation

Distillation uses teacher reasoning to build training examples

A teacher model generates an answer and reasoning trace; the verbatim trace is then reviewed for use in training.

FULL QUESTION A frameshift mutation is a deletion or insertion of one or more of what that changes the reading frame of the base sequence? **Make sure to put the final answer (and only the final answer).**

TEACHER MODEL REASONING / VERBATIM TRACE

We need answer user's question: "A frameshift mutation is a deletion or insertion of one or more of what that changes the reading frame of the base sequence? Make sure to put the final answer (and only the final answer)."

Need final only. The answer: nucleotides? or bases? Let's parse. A frameshift mutation is a deletion or insertion of one or more nucleotides (or bases) in DNA/RNA sequence that changes reading frame. Question says "one or more of what that changes the reading frame of the base sequence?" likely "nucleotides" or "bases". If asking "of one or more of what" maybe from definition: "insertion or deletion of one or more nucleotides". Since they mention base sequence, maybe "nucleotides". Need only final answer. Could be "nucleotides". Ensure only final answer.

DISTILLED ANSWER

Nucleotides

Only the final answer

as instructed

Distillation becomes more demanding as the reasoning becomes more applied

Power systems questions often combine multiple disciplines, assumptions, and sources of evidence.

Basic science: easy to distill

“A frameshift mutation is a deletion or insertion of one or more of what?” A settled fact, a short trace, and an answer any textbook can confirm.

Why it works

The answer is closed-form, so a teacher trace can be accepted or rejected on the spot

What it takes

one settled fact • one answer • a fast reference check

Applied science: power systems

“Should a 500-kW inverter (capable of operating at ± 0.9 pf) absorb reactive power at a 1.06 pu PCC voltage, and is a settings change cheaper than new equipment?” One answer, five disciplines.

Why it is demanding

No single reference settles it, so both the answer and the teacher's trace need expert verification

What it takes

domain SMEs • simulation • cost and standards data

When does validating the reasoning require more effort than validating the final answer?

Sources

Preparation

Q&A Design

Supervision

Adaptation

Evaluation

Useful supervision requires more than plausible reasoning

In applied domains, the final answer may look reasonable while the reasoning still needs validation.

Two traces → different reasoning quality → different training value

Trace A: more suitable for supervision

Question: Why do digital communication links make the traditional concept of burden largely obsolete?

“In a conventional instrument-transformer circuit, relays, meters, and wiring electrically load the secondary. That burden can affect measurement accuracy. With digital sampled values, downstream devices receive data rather than drawing power from the measurement circuit. The variable downstream burden is therefore largely eliminated.”

✓ Grounded ✓ Direct ✓ Few assumptions

Trace B: needs additional validation

Question: Why can an MOCT sensor head connect directly to a 230-kV bus while the fiber path requires insulation as it transitions toward ground?

“The sensor head is likely at the same potential as the bus, so perhaps it does not require a separate insulator. The fiber cable may need a free-standing insulator because surface tracking, flashover, mechanical support, or voltage grading could become concerns as it approaches ground potential. The insulator probably provides the needed electrical and mechanical transition.”

⚠ Speculative ⚠ Multiple hypotheses ⚠ Requires validation

Sources

Preparation

Q&A Design

Supervision

Adaptation

Evaluation

Different adaptation methods change different things

Different methods target different capabilities and require different checks.

CPT

Continued Pre-Training

Continue pretraining on domain corpus

Primary Objective

Domain knowledge acquisition

Check for

General capability retention

SFT

Supervised Fine-Tuning

Curated demonstrations

Primary Objective

Task behavior and response format

Check for

Breadth and reasoning quality

RL/post-training

Reinforcement Learning

Optimize against reward or feedback

Primary Objective

Optimize desired behaviors

Check for

Reward alignment and unintended behavior

SDFT

Self-Distillation Fine-Tuning

Generated demonstrations with filtering / retention aim

Primary Objective

Capability acquisition with stronger retention

Check for

Demonstration quality and retention evidence

For every adaptation step, measure both the intended gain and what remained stable

Sources

Preparation

Q&A Design

Supervision

Adaptation

Evaluation

Domain adaptation may result in some gains while incurring some losses

Gains should improve the target capability while preserving broader model behavior.

Before adaptation

Adaptation

After adaptation

Adaptation success

=

Domain gain

+

Capability retention

Target capabilities

(intended gains)

- Domain knowledge
- Domain task performance
- Response discipline

Capabilities to retain

- General reasoning
- Instruction following
- Uncertainty & refusal behavior
- Out-of-domain knowledge
- Safety & guardrails

Sources

Preparation

Q&A Design

Supervision

Adaptation

Evaluation

Training should teach. Evaluation should show what transferred.

Evidence is needed that the model learned the material and can apply it to new problems.

Teach

Training / development data

Build:

- ✓ Domain knowledge
- ✓ Task behavior
- ✓ Targeted capabilities

Check learning

Held-out, corpus-linked questions

Did the model learn the knowledge or capabilities introduced during adaptation?

Check transfer

Held-out, broader domain questions

Can the model apply what it learned to new power-sector problems?

The stronger evidence is transfer: applying what was learned to problems that the model did not train on.

Sources

Preparation

Q&A Design

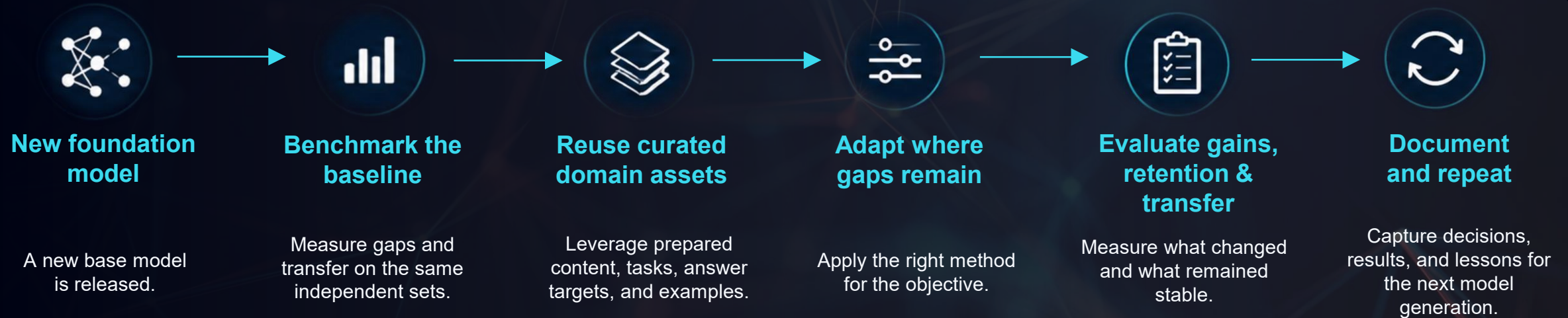
Supervision

Adaptation

Evaluation

The process should be repeatable as foundation models evolve

As foundation models improve, the same process can show where adaptation still adds value.



If we can't repeat the process with the next foundation model, we haven't built something that will scale.

Five questions to explore together

Where shared experience can sharpen how we build and evaluate domain models

- 01 | What makes technical content **training-ready**?
- 02 | How do we create hard questions without accidental **ambiguity**?
- 03 | What makes a **reasoning trace** worth training on?
- 04 | How **separate** should training and evaluation data be?
- 05 | How do we show domain **gains without losing** broader capability?

The five questions point to three areas for deeper work

Where would deeper collaboration be most useful?



Data & task quality

Preparing content, designing questions, defining answer targets



Adaptation

Choosing methods and preserving broader capability



Evidence

Separating evaluation, testing transfer, understanding failure modes

Domain Intelligence Working group

**Future Discussion Topics – Please
sign up**

<https://forms.office.com/r/AfcAdPDWgX>





TOGETHER...SHAPING THE FUTURE OF ENERGY®