Agentic AI for Utility Operators

Real power-sector uses, engineering tools and measured research results

90 minutes / two worked exercises

Alaa Selim
MIT/Dartmouth

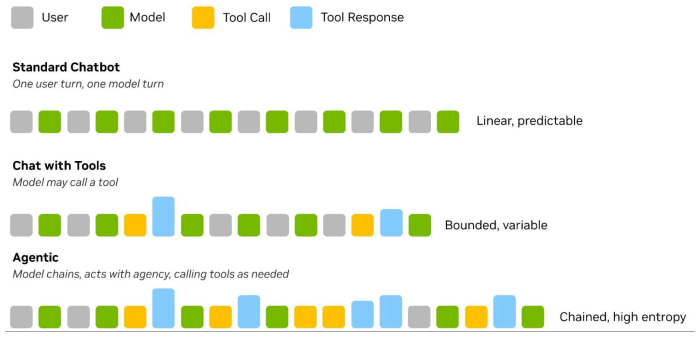
Mohamed Chehada
UT Austin

IEEE ComSoc IEEE Communications Society | IEEE

Independent educational material / IEEE ComSoc theme adaptation

Source: IEEE Communications Society. IEEE Communications Society: official resources and network presentation template. Reviewed 17 October 2026.

What is agentic AI?

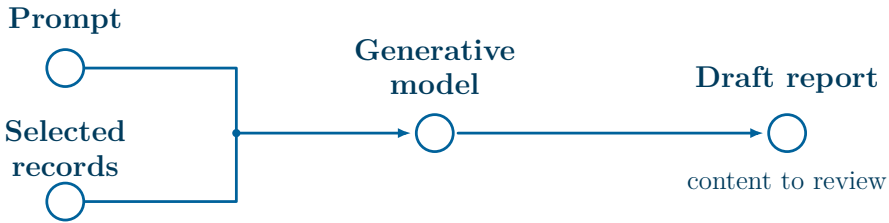


Original NVIDIA interaction patterns. The same generative model can be part of a chat or an agentic system.

- An agentic system uses feedback to choose and adjust actions toward a goal. One agent may be enough.

Source: Erik S., Barry Zhang. Building effective agents. Anthropic Engineering. 2024-12-19.

Generative AI creates content

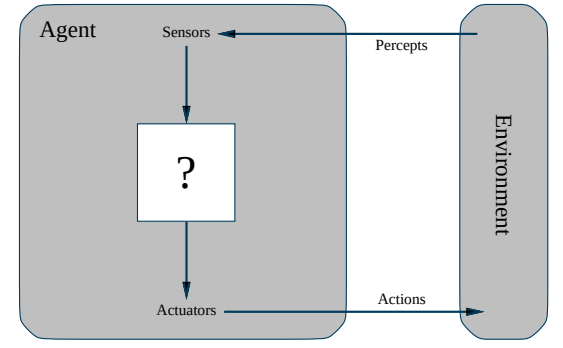


Illustrative report-drafting example. Generative models can also be components of agents.

- Generative AI makes text, images and other content. Large language models (LLMs) are used for text.
- A single call can create a draft. An agent system can use the same model to choose steps.

Source: National Institute of Standards and Technology. Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. NIST AI 600-1. July 2024.

What is an AI agent?

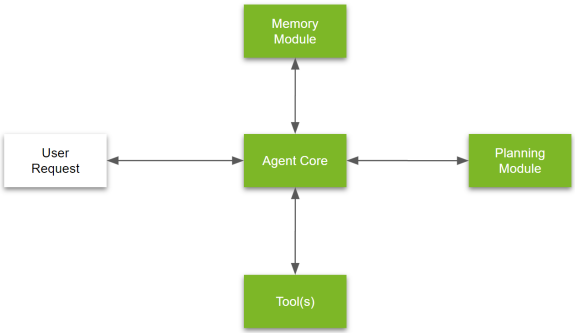


Russell and Norvig's original agent diagram, Figure 2.1. A software agent's sensors and actuators can be software interfaces.

- An agent observes an environment and chooses actions that affect it.
- In a study, an engineering solver calculates grid behavior. The agent reads its results and chooses the next tool call.

Source: Stuart J. Russell, Peter Norvig. Artificial Intelligence: A Modern Approach, Chapter 2: Intelligent Agents. Prentice Hall, second edition, Chapter 2, pp. 32–58. 2003.

An agent chooses a step. A tool runs it.



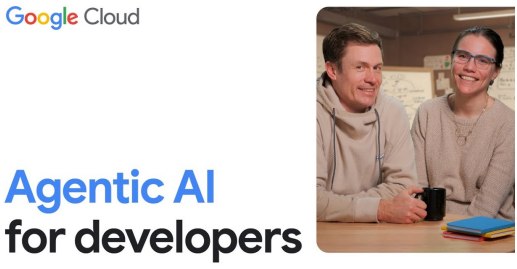
Original NVIDIA agent design: the core connects a request with tools, planning and memory. Other designs can combine these roles.

- The agent chooses an allowed step. A tool carries out the request.
- An engineering solver calculates grid behavior. The agent uses the returned evidence.

Source: Tanay Varshney. Introduction to LLM Agents. NVIDIA Technical Blog. November 30, 2023.

Watch: how an agent chooses its next step

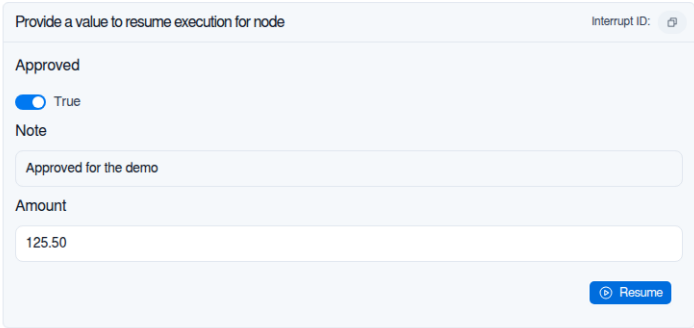
Concept video / Google Cloud Tech / 0:43–3:28



- An agent uses tool results to choose what to do next.
- For our operator: study request to solver to result checks to human review.

Source: Aja Hammerly, Jason Davenport. Agentic AI: Workflows vs. agents. Google Cloud Tech / YouTube. 19 March 2025.

An agent needs a clear goal and limits

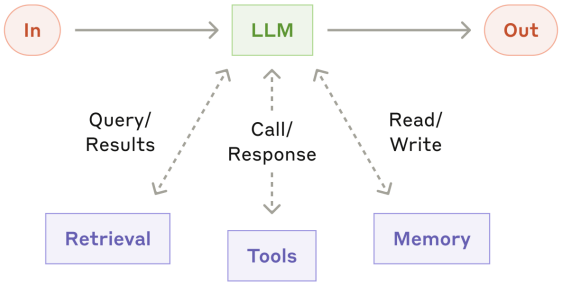


Original LangGraph Studio review-form excerpt. This is a generic demo, not a utility interface.

- The agent can run agreed studies and prepare a report with evidence.
- The operator reviews the findings and decides whether an operating action is allowed.

Source: LangChain documentation maintainers. LangGraph Interrupts. Official LangChain documentation. Accessed 2026-10-17; live documentation.

A model can use records, tools and memory



Original Anthropic building-block figure. It is a general pattern, not a diagram of PowerAgent or PowerDAG.

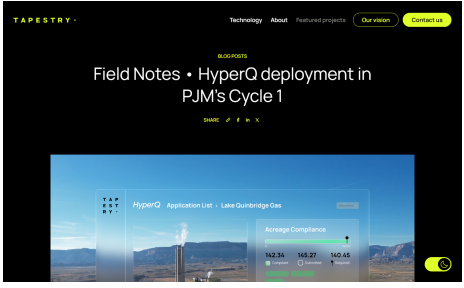
- Retrieval finds records; tools run tasks; memory keeps useful context.
- LangGraph can manage the steps. PowerAgent supplies power-domain components; PowerDAG is one specific system.

Source: Erik S., Barry Zhang. Building effective agents. Anthropic Engineering. 2024-12-19.

PJM uses HyperQ to review connection applications

United States / Supplier reports deployment / PJM confirms earlier testing

- Tapestry's June 2026 report says HyperQ entered PJM Cycle 1 on 29 April. PJM confirmed testing in March.
- Gemini agents check site-control agreements (rights to use land), property parcels and tariff evidence. Engineers make final decisions.
- This work checks application readiness before the electrical study begins.

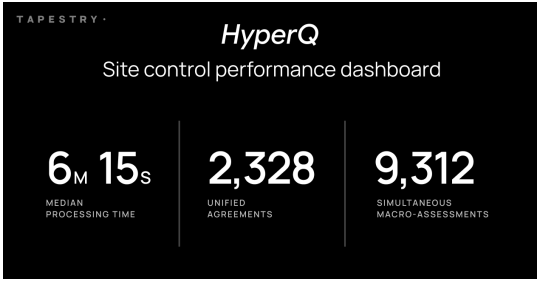


Original Tapestry article and published animation still. Supplier report, June 2026.

A numerical study needs reliable input evidence.

Source: Tapestry. Field Notes / HyperQ deployment in PJM's Cycle 1. 2026-06-11; deployment entry 2026-04-29.

HyperQ: what the reported numbers measure



Original Tapestry performance graphic, June 2026. The timing covers only computer processing.

- The timings cover computer processing for 811 applications. They exclude document-transfer time and human validation.

Source: Tapestry. Field Notes / HyperQ deployment in PJM's Cycle 1. 2026-06-11; deployment entry 2026-04-29.

CAISO/OATI: agents help review planned outages

- OATI's AI Genie uses Workflow, Advisor and Recon agents under one coordinator. Operators keep final authority.
- The supplier reports processing 500+ next-day requests nightly in under eight minutes. It gives no per-request time or reviewer-effort result.



Real California ISO control room. OATI supplies the batch result and later production-use claim.

A processing rate needs a clear work boundary and a measure of the review left for people.

Source: OATI. Generative and Agentic AI in Grid Operations: Field Lessons from OATI, CAISO, PNNL, and EPRI. 2026 (exact publication day not displayed).

Duke and AWS: agents help prepare planning studies

- Duke engineers use AWS-managed agents to prepare study cases and coordinate interconnection studies for new grid connections.
- A September 2026 AWS announcement reports data preparation falling from two weeks of manual work to hours.
- The final number of hours is unspecified. Engineers keep final responsibility for engineering decisions.



Duke operator, 2022, shown for context. The reported agents support planning, rather than this pictured operating desk.

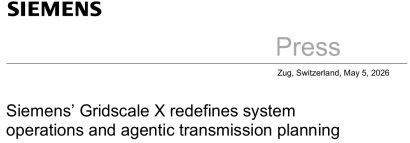
Faster preparation alone does not establish a shorter full regulatory process.

Source: Amazon Web Services. AWS Launches Agentic Grid Planning Program to Accelerate Interconnection Studies. 2026-09-17.

Siemens announces agents for PSS E planning studies

Product announcement / 5 May 2026

- Siemens' Gridscale X announcement describes agents coordinating transmission-planning simulations.
- The targets include connection studies for large loads and data centers, using open software interfaces (APIs).
- The announcement does not name a production user of the new agentic PSS E features.


SIEMENS

Press

Zug, Switzerland, May 5, 2026

Siemens' Gridscale X redefines system operations and agentic transmission planning

Original Siemens press-release header, 2026. This is an announced product capability.

Check product features and measured utility results separately.

Source: Siemens. Siemens' Gridscale X redefines system operations and agentic transmission planning. 2026-05-05.

National Grid reports agents checking project plans

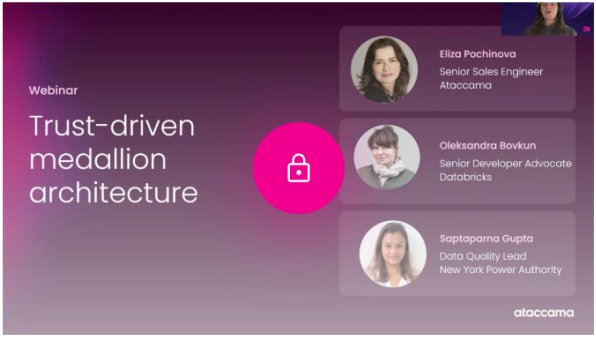
Reported current use / UK–US group / Rollout location unspecified

- The results transcript says its "capital control tower" already checks early project options and documents at approval stages.
- Project managers get feedback while they can still change their plans.
- The group reports no agent framework, timing result or accuracy result.

Source: National Grid. National Grid Full Year Results 2025/26: Video Webcast Transcript. 2026-05-14.

NYPA: better data for agent use

NYPA data-quality lead / Webinar, 14 October 2026



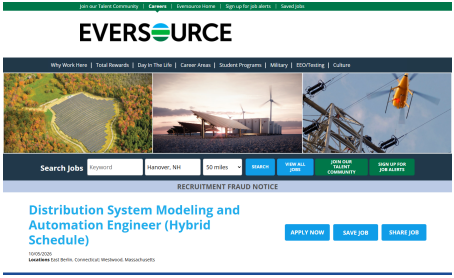
Original Ataccama webinar artwork. The NYPA speaker appears at lower right.

- NYPA describes checks on meter data before AI uses it.
- A separate MCP demo uses product data. It does not show NYPA grid control.

Source: Session speakers: Saptaparna Gupta (NYPA), Oleksandra Bovkun (Databricks), Eliza Pochinova (Ataccama). On demand: Trust-driven medallion architecture: How to make your CDEs ready for AI. Ataccama. 2026.

Eversource: adding agentic skills to grid studies

Eversource / Existing simulation tools and a stated engineering direction



Original Eversource vacancy, 5 October 2026.

- Eversource uses MATLAB and Synergi Electric for distribution-network studies.
- Its October 2026 engineering vacancy asks for agentic and AI-based workflows in MATLAB. The vacancy shows hiring intent, not deployment.

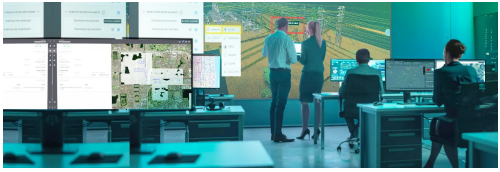
Integration of agentic and AI-based workflows in MATLAB

Original requirement excerpt. Hiring and integration intent; deployment is not established.

Source: Eversource Energy (employer). Distribution System Modeling and Automation Engineer (Hybrid Schedule). Eversource Energy. 2026-10-05.

GE Vernova: a platform that can connect grid tools

- Alabama Power's use of GridOS for Distribution gives a named example of GE Vernova's grid platform.
- GridOS provides shared software for coordinating grid work. This shows where study tools and operator workflows could connect.
- The public statement does not separate out a reliability improvement caused by agents.



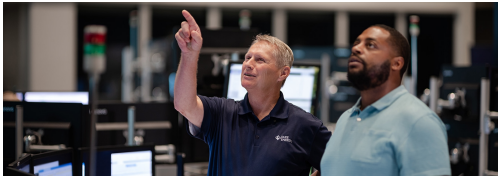
Official GridOS promotional composite. The platform example does not establish a measured agent-specific result.

For each case, identify the customer, task, agent's role and measured result.

Source: GE Vernova. GE Vernova Launches GridOS for Distribution, the Industry's First Unified Solution for Grid Orchestration. 2026-02-03.

Operators often work against a deadline

- A PNNL survey of 20 control-room operators reported that, on average, 61% of tasks needed completion in under 30 minutes.
- The survey had 75 questions. It gives no task count or statistical spread behind that average.
- An assistant should reduce the time people spend finding and reviewing evidence within the task's deadline.



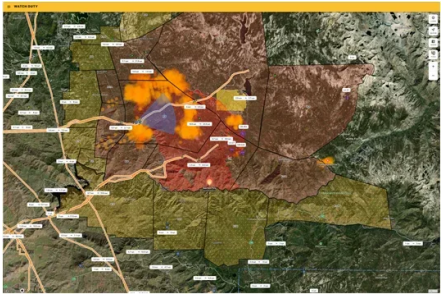
Duke Energy control room, 2022. The PNNL survey is a separate study.

Measure the operator's total effort and deadline alongside AI response time.

Source: Corey Fallon, Brett Jefferson, Tyler Ryan. An Analysis of Grid Operators. Pacific Northwest National Laboratory. 2023-04-19.

Three checks for a useful recommendation

- Data: do the asset, units, time and grid connections (topology) match the question?
- Engineering: does the solution meet the requested equations and limits?
- Decision: is there enough evidence for the next step the user is allowed to take?



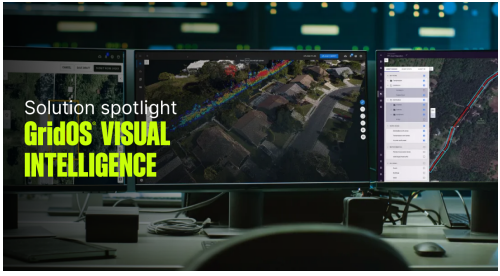
Real CAISO Watch Duty display. It combines fire and transmission-line information for operator awareness.

Agree on acceptance checks for the whole task.

Source: John Phipps. Control center modernization is enhancing reliability. California ISO. 2025-09-10.

A clear task makes success easier to test

- The task is to check whether an earlier study result still applies after a new event.
- Return evidence or ask for missing data. Stop if inputs change, a tool fails, the work budget runs out or permission is missing.

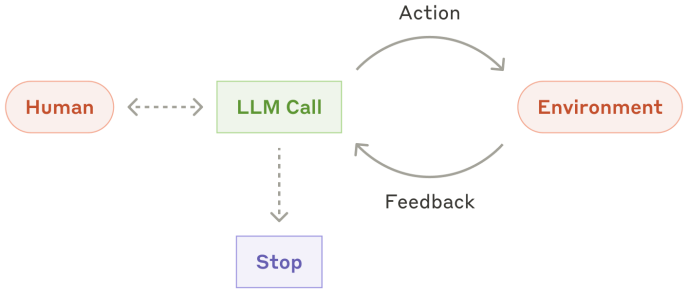


GE Vernova's published GridOS Visual Intelligence screen. This product image supplies task context, not an agent result.

A clear task gives us a clear way to check success.

Source: GE Vernova. GridOS Visual Intelligence product screens. Accessed 17 October 2026.

An LLM agent acts, reads the result and decides again

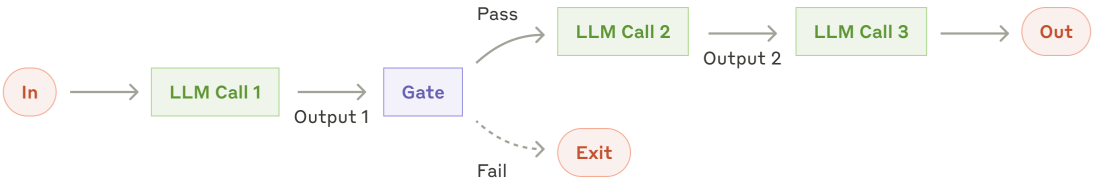


Original Anthropic agent-loop figure. In a grid study, feedback can be solver results or a failed check.

- After an action, the agent reads what happened before choosing the next step.
- It can call another allowed tool, ask a person or stop. An error is also useful feedback.

Source: Erik S., Barry Zhang. Building effective agents. Anthropic Engineering. 2024-12-19.

When data are stale, the next step is to stop



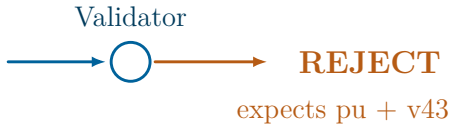
Original Anthropic gated workflow. In our study example, a stale input should fail the check and stop the sequence.

- The asset lookup succeeds, but the measurement request marks the data as stale, meaning out of date.
- Ask for current evidence. The original task remains unresolved.

Source: Erik S., Barry Zhang. Building effective agents. Anthropic Engineering. 2024-12-19.

Unit checks catch a kV versus p.u. mistake

```
asset=F12, voltage=12.47 kV
topology=v42
```




```
asset=F12, voltage=1.00 pu
base=12.47 kV, topology=v43
```

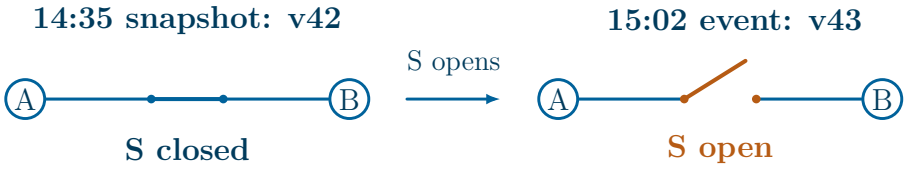


12.47 kV / 12.47 kV = 1.00 pu, using the same voltage base

Teaching example with invented values. These values are not recommended operating limits.

- 12.47 kV is 1.00 per unit (p.u.) only with a 12.47 kV base. Per unit means value divided by base.
- A tool expecting p.u. must reject a raw kV value or require explicit information for converting it.

A study must match the current grid connections

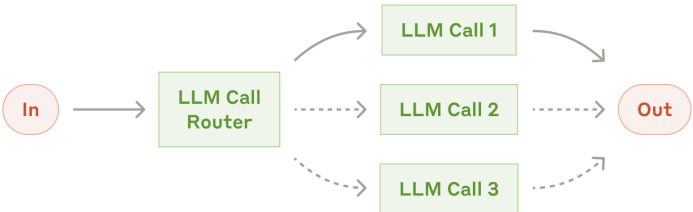


Old result needs revalidation against the changed topology

Invented before-and-after network. This illustration does not calculate power flow.

- The old study used a closed switch and grid version v42. The event opens the switch and creates version v43.
- Update the inputs and repeat the affected calculations. Rewriting the explanation does not update the numerical result.

A solver result depends on the problem it solves



Original Anthropic routing pattern. Choosing an engineering study also requires selecting the right problem and tool.

- Choose the study first: power flow, optimal dispatch and dynamic simulation answer different questions.
- A solver's success flag covers that problem. Also check the requested cases, grid limits and permission to act.

Source: Samim Konjicija, Slaven Peles. Integrating Agentic Artificial Intelligence with High-Performance Computing for Grid Planning. arXiv; EnergyCon 2026 proceedings forthcoming as described by authors. 3 September 2026; v1.

Exercise 1: check three returned data packets

Work in pairs / 5 minutes, then 1-minute answer

```
A = {age_s:20, topology:v42, unit:pu}
B = {age_s:15, topology:v43, unit:kV}
C = {age_s:45, topology:v43, unit:pu}
```

- For this exercise, data must be at most 60 s old, use topology v43 and use p.u. The user can recommend actions only.
- Which packet passes? Explain why each other packet fails.
- Does passing these checks establish complete engineering validity or permission to execute an action?

Exercise 1 answer: C passes the first checks

Worked answer / data checks

A: 20 s / v42 / p.u.	check →	Stop: v42 is the wrong topology
B: 15 s / v43 / kV	check →	Stop: kV is the wrong unit
C: 45 s / v43 / p.u.	check →	**Pass: these initial checks only**

The 60-second limit and packet labels belong to this teaching exercise.

- C is recent enough, uses v43 and has p.u. units. A has the wrong topology; B has the wrong unit.
- Passing these checks is only a start. Check the model and engineering limits. The user may recommend, not execute.

Power-system papers test different tasks

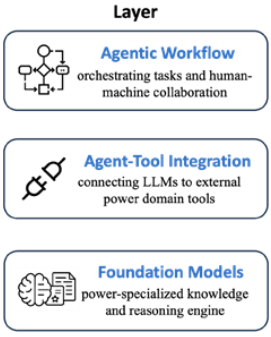
- Fault-analysis tools find events in recorded waveforms. Simulation assistants set up and run studies.
- Control prototypes suggest actions in model grids. Benchmarks test how well systems find evidence or answer numerical questions.
- Each result applies to the tested task, tools, allowed actions and test cases.



Image accompanying the EPRI Open Power AI Consortium announcement, published by NVIDIA, 20 March 2025.

Source: Barun Kumar Saha, Aarthi V, O. D. Naidu. DrAgent: An Agentic Approach to Fault Analysis in Power Grids Using Large Language Models. 2025 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), pp. 938–945. 2025-02-18–21.

Harvard PowerAgent: models, tools and study workflows

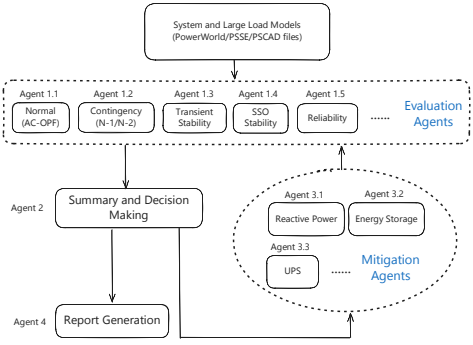


Original Harvard author roadmap: workflows, tool connections and foundation models. Three-layer excerpt.

- PowerAgent's 2025 IEEE roadmap links AI models, power tools and study workflows.
- PowerFM supports modeling. PowerMCP connects tools. PowerWF manages workflows. PowerAgentBench-Dyn is a later benchmark, not the whole initiative.

Source: Qian Zhang, Le Xie. PowerAgent: A Road Map Toward Agentic Intelligence in Power Systems: Foundation Model, Model Context Protocol, and Workflow. IEEE Power and Energy Magazine, volume 23, issue 5, pages 93–101. 2025.

PowerAgent example: what changes when load is added?

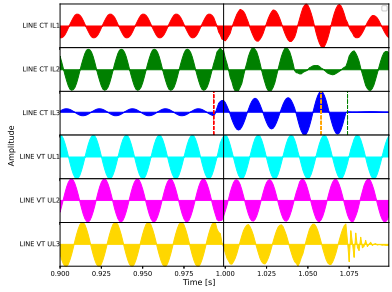


Original Harvard concept for grid-impact studies. The separate PowerWF example uses LangGraph and PowerWorld; this is not its exact graph.

- It runs power flow and single-outage checks (N-1) for the original case and the added-load case.
- LangGraph manages the steps and saved results. PowerWorld calculates the electrical results.

Source: PowerAgent / PowerWF community contributors. PowerWF Grid Impact Evaluation — LangGraph workflow source. Author-maintained source code, Grid Impact Evaluation/src/agent/graph.py. Live source inspected 17 October 2026.

DrAgent: reading a recorded grid fault

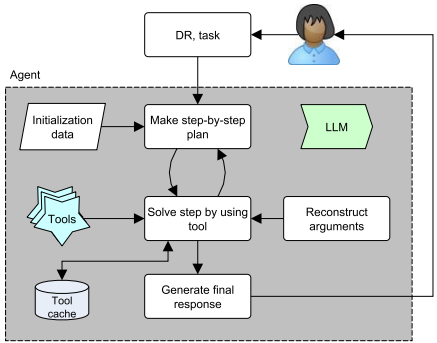


Original DrAgent Figure 2, PDF p.4. The third current channel shows a fault in a sample disturbance record.

- The third current trace shows the fault. Dashed lines mark its start, highest current and clearance (when the fault ends).

Source: Barun Kumar Saha, Aarthi V, O. D. Naidu. DrAgent: An Agentic Approach to Fault Analysis in Power Grids Using Large Language Models. 2025 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), pp. 938–945. 2025-02-18–21.

DrAgent selects tools to answer a fault question

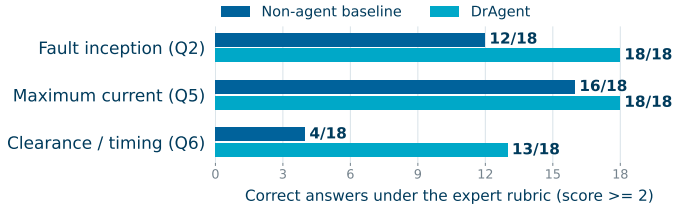


Original DrAgent architecture: planning, tool execution and answer generation.

- The plan chooses signal-analysis tools or manual search to match the question.
- Signal tools measure the fault. Manual search finds product information in documents.

Source: Barun Kumar Saha, Aarthi V, O. D. Naidu. DrAgent: An Agentic Approach to Fault Analysis in Power Grids Using Large Language Models. 2025 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), pp. 938–945. 2025-02-18–21.

DrAgent: more correct answers, with errors still present

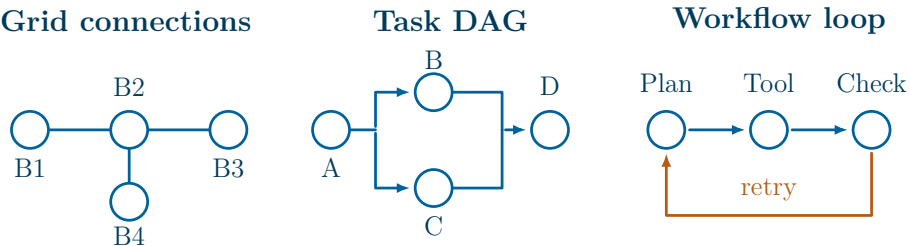


Counts reconstructed from Table IV and the Appendix. They are neither new experiments nor the paper's mean-score measure.

- We counted scores of 2 or more as correct across the paper's 18 recorded cases.
- Timing the fault's end remains harder. The grading rules also accept a correctly labeled fault duration.

Source: Barun Kumar Saha, Aarthi V, O. D. Naidu. DrAgent: An Agentic Approach to Fault Analysis in Power Grids Using Large Language Models. 2025 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), pp. 938–945. 2025-02-18–21.

Three kinds of graph have different jobs

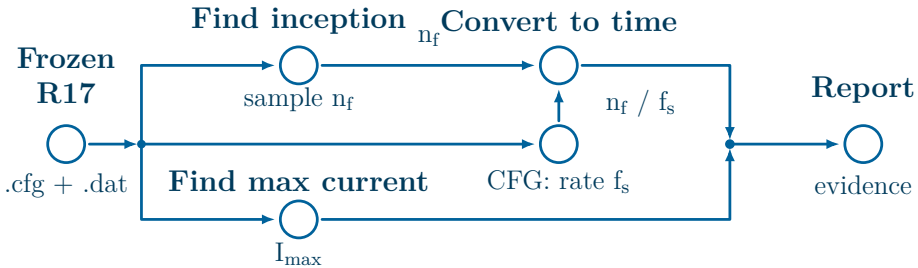


Three different meanings of "graph." These small examples explain connections, not calculated electrical results.

- Grid edges show electrical connections. A task DAG (directed acyclic graph) links calculations through the results they need.
- Workflow edges show which software step can run next, including a retry loop.

Source: Barun Kumar Saha, Aarthi V, O. D. Naidu. DrAgent: An Agentic Approach to Fault Analysis in Power Grids Using Large Language Models. 2025 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), pp. 938–945. 2025-02-18–21.

A DAG shows which result must come first

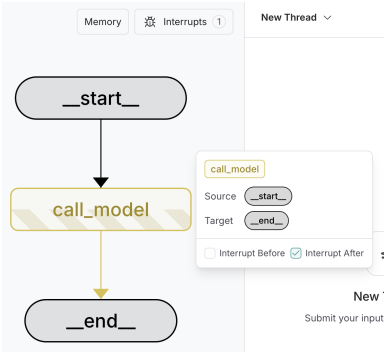


Original teaching adaptation of DrAgent's tool dependencies, rather than its exact execution graph.

- A DAG links results to calculations that need them. Its arrows do not form a loop.
- Run independent calculations when inputs are ready. Wait for every required result before reporting.

Source: Barun Kumar Saha, Aarthi V, O. D. Naidu. DrAgent: An Agentic Approach to Fault Analysis in Power Grids Using Large Language Models. 2025 International Conference on Artificial Intelligence in Information and Communication (ICAIIC), pp. 938–945. 2025-02-18–21.

LangGraph can pause to inspect a run



Real LangGraph debugger screenshot. A paused run can be inspected before the next software step.

- A debugger can pause the graph so a developer can inspect it. A failed check can also send execution back to an earlier step.
- Software routing does not check the grid model. The application still checks inputs and engineering results.

Source: LangChain documentation maintainers. LangGraph Graph API overview. Official LangChain documentation. Accessed 2026-10-17; live documentation.

LangGraph demo: pause, inspect and continue

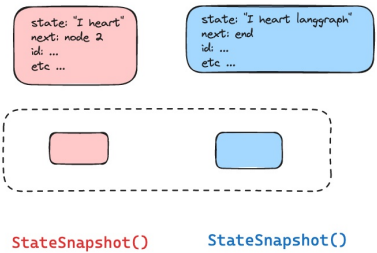
Official LangChain demo / 99-second excerpt / 3:34–5:13



- A graph can pause so its state and outputs can be inspected.
- For a grid study, check the solver result before continuing.

Source: LangChain. Strategies for debugging agents with LangGraph Studio. LangChain / YouTube. 26 March 2025.

Saved state must identify the study and its evidence

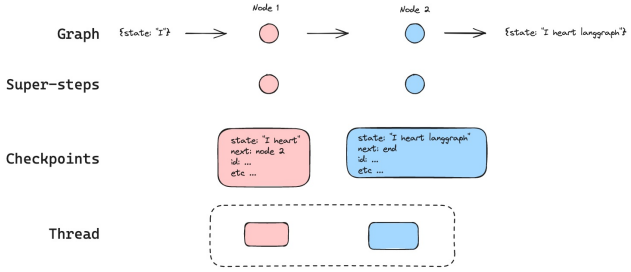


Original LangGraph state diagram. Software checkpoints still need grid-case, time and evidence identifiers.

- Save the case ID, input versions, evidence, status and attempt count.
- Label each result with its case or outage ID. This helps reveal missing or duplicate results when parallel work comes together.

Source: LangChain documentation maintainers. LangGraph Graph API overview. Official LangChain documentation. Accessed 2026-10-17; live documentation.

A checkpoint lets a workflow pause and resume

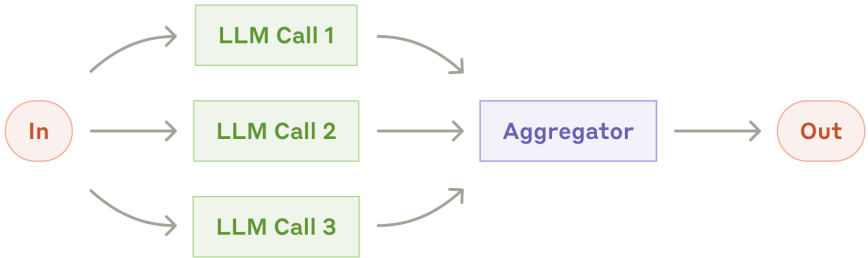


Original LangGraph checkpoint diagram. Saving progress does not prove that the grid data still match.

- A checkpoint saves progress; thread_id links the saved steps to the same workflow.
- Changed grid connections can make old results invalid. Check them and rerun affected studies.

Source: LangChain documentation maintainers. LangGraph Persistence. Official LangChain documentation. Accessed 2026-10-17; live documentation.

Parallel outage studies must account for every case

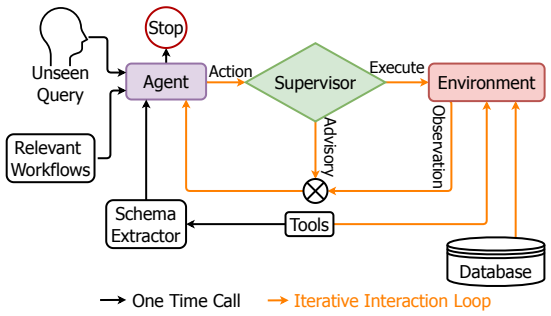


Original Anthropic parallel workflow. A grid study must also account for missing or failed outage cases.

- Run each selected outage on a separate, consistent copy of the same starting grid case.
- Before combining results, compare the returned case IDs with the expected list. Mark failed or missing cases as unresolved.

Source: LangChain documentation maintainers. LangGraph Graph API overview. Official LangChain documentation. Accessed 2026-10-17; live documentation.

PowerDAG checks the steps needed before a tool runs

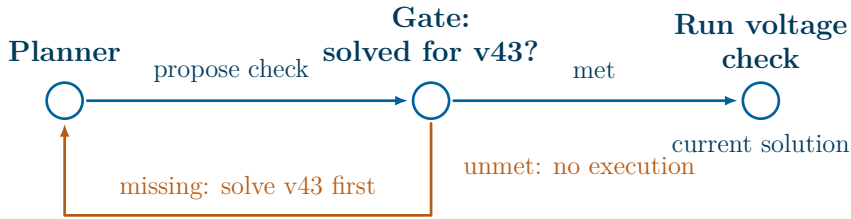


Original PowerDAG v4 architecture, including the supervisor before environment execution.

- PowerDAG finds relevant verified workflow examples. It uses a dependency graph and a gate that checks required earlier tool calls.
- The benchmark covers 108 tools, ten task families and four Vermont feeder models (local distribution networks).

Source: Emmanuel O. Badmus, Amritanshu Pandey. PowerDAG: Supervisory Agentic AI System for Automating Distribution Grid Analysis. arXiv preprint, version 4. 5 July 2026 (v4); first submitted 18 March 2026.

PowerDAG: solve the grid before checking its voltage

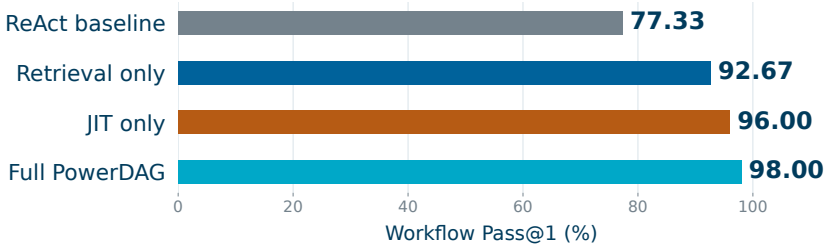


Teaching adaptation: the current-v43 check adds a freshness rule. The paper's gate checks which required tools have already run.

- In the paper's example, a voltage-violation check needs a solved network state first.
- If an earlier required step is missing, the gate rejects the call and tells the agent what to do first.

Source: Emmanuel O. Badmus, Amritanshu Pandey. PowerDAG: Supervisory Agentic AI System for Automating Distribution Grid Analysis. arXiv preprint, version 4. 5 July 2026 (v4); first submitted 18 March 2026.

PowerDAG results: the same queries, four setups



PowerDAG v4, Table II. ReAct: a reason-and-act loop. JIT: prerequisite checks just before a call.

- Each GPT-5.5 setup runs the same 150 test queries once.
- Pass@1 measures first-run workflow success: required tool calls, valid inputs and correct order. It does not measure field reliability.

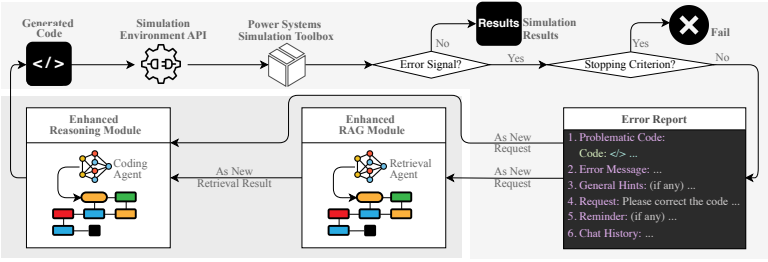
Source: Emmanuel O. Badmus, Amritanshu Pandey. PowerDAG: Supervisory Agentic AI System for Automating Distribution Grid Analysis. arXiv preprint, version 4. 5 July 2026 (v4); first submitted 18 March 2026.

PowerDAG: what can others check and reuse?

- The public benchmark provides tasks, tool descriptions, workflow records and scoring code.
- Three task groups can run with the public files. Seven need private grid-map data and network tools.
- Reuse the idea of checking required earlier steps. Test the new tools, case state and engineering results.

Source: Emmanuel O. Badmus and contributors. DistGrid-AgentBench: A Distribution Grid Benchmark for Agentic AI. Author-maintained GitHub repository. Live repository reviewed 17 October 2026.

Simulation agents find functions, run code and repair errors

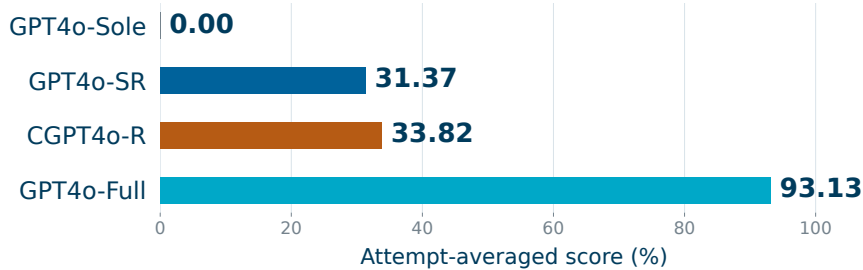


Original environmental acting module: simulation errors feed retrieval and code repair.

- The agent finds MATLAB functions and settings, writes code and runs the study.
- A software error triggers repair. Code that runs can still use the wrong settings.

Source: Mengshuo Jia, Zeyu Cui, Gabriela Hug. Enhancing LLMs for Power System Simulations: A Feedback-Driven Multi-Agent Framework. IEEE Transactions on Smart Grid, volume 16, issue 6, pages 5556–5572. 2025.

Simulation-agent results: points, not a pass rate

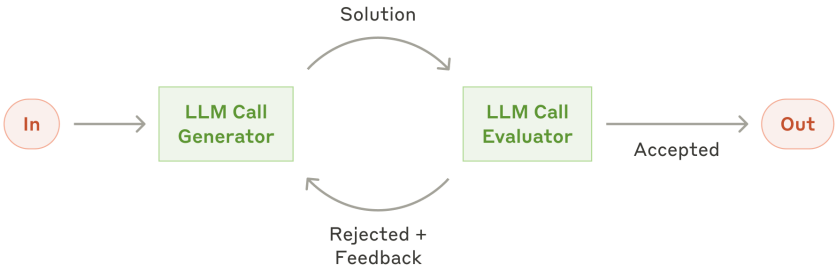


DALINE comparison reported in the author preprint, Figure 8 / section V. Each task allows up to three attempts.

- Across 34 DALINE tasks, the score averages graded attempts and penalizes extra settings. The 93.13 score is not a pass rate.

Source: Mengshuo Jia, Zeyu Cui, Gabriela Hug. Enhancing LLMs for Power System Simulations: A Feedback-Driven Multi-Agent Framework. IEEE Transactions on Smart Grid, volume 16, issue 6, pages 5556–5572. 2025.

Trying again can repeat the same software error



Original Anthropic feedback pattern. A retry must actually use the repaired code and changed inputs.

- R02 found that an agent request and its cache key both left out reactive load.
- Fixing the request still returned the old result: the cache key had not changed.

Source: Eve Tsybina, Samim Konjicija, Slaven Peles. Agentic Artificial Intelligence for Power Systems: Strategies to Identify and Close Capability Gaps. arXiv. 30 July 2026; v1.

Siemens PSS E: agentic transmission planning

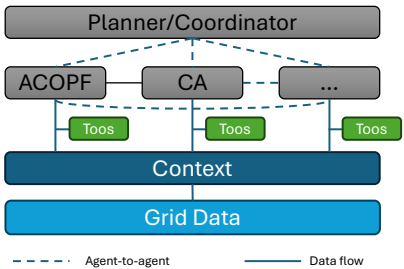
Official grid-platform video / 79 seconds



- AI helps prepare and review studies; PSS E supplies the grid simulation.
- Connect to our lab: define the case to run the solver to check to review.

Source: Siemens. Gridscale X PSS E: agentic transmission planning — product video. Undated video; accessed 26 October 2026.

GridMind 2025 connects typed agents to solvers

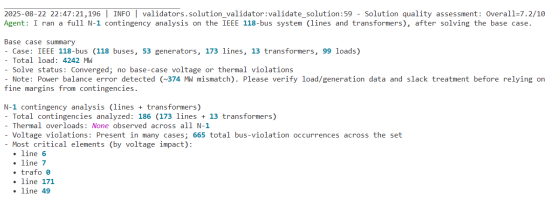


Original GridMind2025 diagram of specialized agents and numerical analysis tools.

- GridMind 2025 uses PydanticAI and pandapower for AC optimal power flow (ACOPF) and outage studies.
- It reports 100% ACOPF success on case 118 across six models. That does not prove complete outage coverage.

Source: Hongwei Jin, Kibaek Kim, Jonghwan Kwon. GridMind: LLMs-Powered Agents for Power System Analysis and Operations. arXiv; published as SC Workshops 2025, DOI 10.1145/3731599.3767409. 2025-09-02 (arXiv v1); SC Workshops 2025.

GridMind: reported answers need independent checks



A separate published response claims no thermal overloads and flags a balance error. It is not independent ground truth.

- On case 118, one model reports 165% maximum overload; five report 137%. They also name different critical lines.
- The table has no complete independent outage reference. Agreement alone cannot tell us which answer is right.

Source: Hongwei Jin, Kibaek Kim, Jonghwan Kwon. GridMind: LLMs-Powered Agents for Power System Analysis and Operations. arXiv; published as SC Workshops 2025, DOI 10.1145/3731599.3767409. 2025-09-02 (arXiv v1); SC Workshops 2025.

GridMind 2026 reuses results when their inputs still match

```
Reported session: seven queries on synthetic Texas 2000
Change cascade fan-out k  to  rerun cascade
Shift forecast by 6 h  to  rerun all five stages
Restore the exact old inputs  to  reuse saved results
```

- LangGraph coordinates five stages: weather, outages, OPF, screening and cascade analysis.
- Input fingerprints identify saved results. A changed forecast makes affected results need a new run.

Reuse a result only when the inputs it depends on still match.

Source: D. Adrian Maldonado, Jonghwan Kwon, Sunash B. Sharma, Hongwei Jin, Stefano Fenu, Emil Constantinescu, David Sehloff, Parfait Gasana, Duane Rudolph Verner, Kibaek Kim. GridMind: An Agentic Workflow for Weather-Driven Power Grid Risk Assessment. KDD 2026 Workshop on Scientific and Societal Agents / OpenReview. August 2026 workshop; reviewed 17 October 2026.

GridMind: which changes need a new run?

Changed input		Affected stages
`k: 50 to 200`	→	**cascade only** reuse the other four stages
`forecast: +6 h`	→	**line-wind, outages, ACOPF, contingency, cascade**

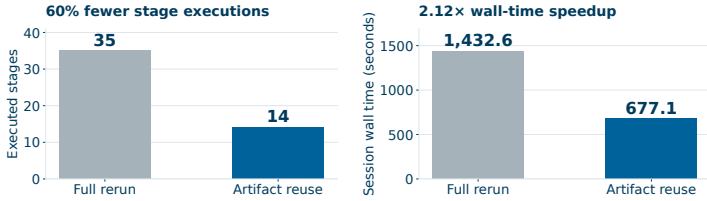
Affected sets from the reported trace; no new dependency edges inferred

Teaching view of GridMind 2026 Table 1. Only the reported affected stages are shown; no extra dependencies are assumed.

- Change only k: rerun cascade. Shift the forecast by six hours: rerun all five stages.
- In this implementation, changing the ACOPF solver reruns ACOPF only.

Source: D. Adrian Maldonado, Jonghwan Kwon, Sunash B. Sharma, Hongwei Jin, Stefano Fenu, Emil Constantinescu, David Sehloff, Parfait Gasana, Duane Rudolph Verner, Kibaek Kim. GridMind: An Agentic Workflow for Weather-Driven Power Grid Risk Assessment. KDD 2026 Workshop on Scientific and Societal Agents / OpenReview. August 2026 workshop; reviewed 17 October 2026.

GridMind results: fewer stages and less time are different



GridMind 2026 Table 1: full reruns in fresh workspaces versus saved-result reuse, on a synthetic 2,000-bus hurricane-risk study.

- For the same seven queries, reuse cuts stage runs by 60% and total run time by 52.7%. Some stages cost more time than others.

Source: D. Adrian Maldonado, Jonghwan Kwon, Sunash B. Sharma, Hongwei Jin, Stefano Fenu, Emil Constantinescu, David Sehloff, Parfait Gasana, Duane Rudolph Verner, Kibaek Kim. GridMind: An Agentic Workflow for Weather-Driven Power Grid Risk Assessment. KDD 2026 Workshop on Scientific and Societal Agents / OpenReview. August 2026 workshop; reviewed 17 October 2026.

Exercise 2: check whether saved weather-risk results still apply

Group exercise / 6 minutes, then 2-minute answer

- A saved run uses forecast F0, topology v43 and cascade setting k=50. Only k changes to 200. What reruns?
- Now shift the forecast by six hours. Which of the five saved stage results still apply?
- A required simulation fails. What status and evidence should the operator receive?

Source: D. Adrian Maldonado, Jonghwan Kwon, Sunash B. Sharma, Hongwei Jin, Stefano Fenu, Emil Constantinescu, David Sehloff, Parfait Gasana, Duane Rudolph Verner, Kibaek Kim. GridMind: An Agentic Workflow for Weather-Driven Power Grid Risk Assessment. KDD 2026 Workshop on Scientific and Societal Agents / OpenReview. August 2026 workshop; reviewed 17 October 2026.

Exercise 2 answer: rerun what changed

Worked answer / cited workflow

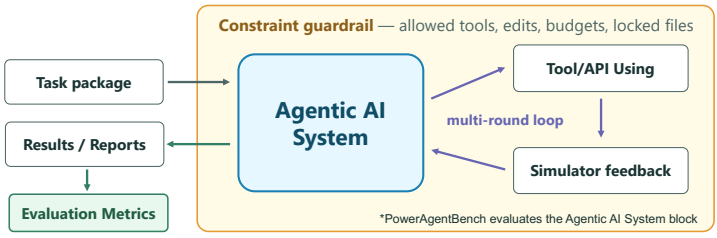
Only k: 50 to 200	→	Rerun cascade; reuse four stages
Forecast shifts by 6 h	→	Rerun all five stages
A required study fails	→	**Return INCOMPLETE, with evidence**

Reuse follows the cited workflow's input dependencies; another implementation can differ.

- A forecast shift invalidates all five saved stages: weather, outages, OPF, screening and cascade.
- If a study fails, return INCOMPLETE with case/stage IDs, input versions, the error and missing conclusions.

Source: D. Adrian Maldonado, Jonghwan Kwon, Sunash B. Sharma, Hongwei Jin, Stefano Fenu, Emil Constantinescu, David Sehloff, Parfait Gasana, Duane Rudolph Verner, Kibaek Kim. GridMind: An Agentic Workflow for Weather-Driven Power Grid Risk Assessment. KDD 2026 Workshop on Scientific and Societal Agents / OpenReview. August 2026 workshop; reviewed 17 October 2026.

PowerAgentBench-Dyn allows only specific model changes

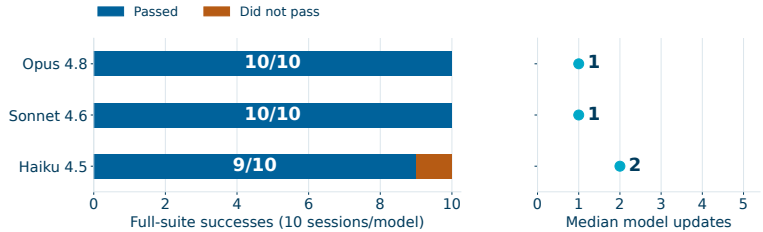


Original benchmark structure: input package, agentic system, tool calls and evaluation metrics.

- The Harvard-led test gives an agent a modified solar-PV dynamic model and feedback from eight DMView/PSS/E quality tests.
- It may change only four REECAU1 controller gains. All other fields are locked; at most five updates are allowed.

Source: Qian Zhang, Andrea Pomarico, Costas Mylonas, Magda Foti, Alberto Berizzi, Le Xie. PowerAgentBench-Dyn: A Benchmark for Agentic AI in Power System Dynamic Studies. arXiv. 2026-06-18 (v1).

PowerAgentBench-Dyn results: passed sessions and updates



PowerAgentBench-Dyn Table II: model-quality task. Other tasks in this benchmark use different models and checks.

- A session passes only if all eight tests pass within five allowed updates. Testing one modified model limits what we can conclude.

Source: Qian Zhang, Andrea Pomarico, Costas Mylonas, Magda Foti, Alberto Berizzi, Le Xie. PowerAgentBench-Dyn: A Benchmark for Agentic AI in Power System Dynamic Studies. arXiv. 2026-06-18 (v1).

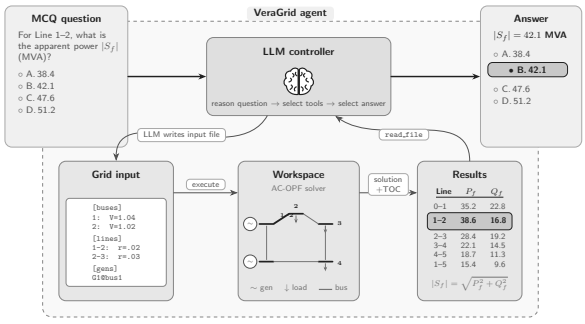
PowerAgentBench-Dyn also checks grid response to faults

- PowerMCP and PowerFactory study ten candidate faults and identify three critical cases.
- Across six models, full-workflow passes range from 0/10 to 10/10. Passing includes correct findings and allowed mitigation.

Different tasks in one benchmark can fail in different ways.

Source: Qian Zhang, Andrea Pomarico, Costas Mylonas, Magda Foti, Alberto Berizzi, Le Xie. PowerAgentBench-Dyn: A Benchmark for Agentic AI in Power System Dynamic Studies. arXiv. 2026-06-18 (v1).

VeraGrid-Agent uses solver results to answer a question

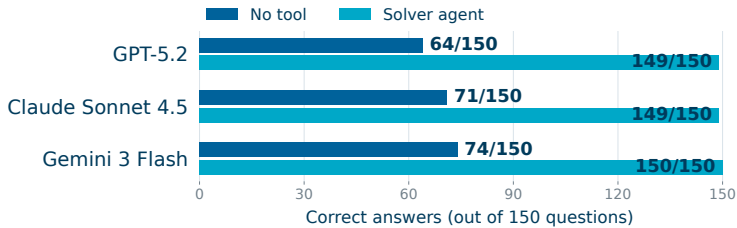


Original VeraGrid-Agent workflow: question, LLM controller, input file, AC-OPF workspace and selected result reads.

- The agent writes feeder inputs, runs VeraGrid AC OPF and reads the relevant result rows.
- It can repair tool errors within a fixed budget, then return a multiple-choice answer.

Source: Shivanshu Tripathi, Hamed Mohsenian-Rad, Maziar Raissi. VeraGrid-Agent: Tool-Augmented LLMs for Distribution Optimal Power Flow at the Grid Edge. arXiv preprint. 2026-07-28 (v1).

VeraGrid-Agent results: the same models with and without tools

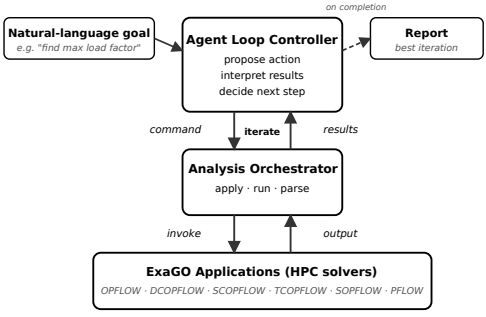


VeraGrid-Agent Table I: selected comparisons using the same model. Missing or badly formatted answers count as incorrect.

- All 150 questions use one fixed feeder. Better multiple-choice answers do not prove general dispatch or field performance.

Source: Shivanshu Tripathi, Hamed Mohsenian-Rad, Maziar Raissi. VeraGrid-Agent: Tool-Augmented LLMs for Distribution Optimal Power Flow at the Grid Edge. arXiv preprint. 2026-07-28 (v1).

AgentiGrid: what does one reported run mean?



Original AgentiGrid data flow: loop controller, analysis-controller interface and ExaGO application.

- The September 2026 paper tests four models on three applications, using one 200-bus system.
- The 12 searches are 12 different setups, not 12 repeats of one task.
- ExaGO does the numerical analysis. The agent chooses changes and explains the results.

Source: Samim Konjicija, Slaven Peles. Integrating Agentic Artificial Intelligence with High-Performance Computing for Grid Planning. arXiv; EnergyCon 2026 proceedings forthcoming as described by authors. 3 September 2026; v1.

A lower cost can come from solving the wrong task

$$\mathcal{U}_{\text{dispatch}} \subseteq \mathcal{U}_{\text{expanded}}$$

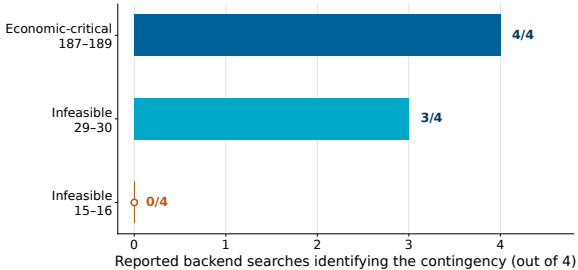
$$\min_{u \in \mathcal{U}_{\text{expanded}}} J(u) \leq \min_{u \in \mathcal{U}_{\text{dispatch}}} J(u)$$

- R01 reports costs of 27,564 before and 17,398 after one model changed unit commitment: which units are on.
- The task allowed only dispatch changes: how much power those units produce. The lower value breaks that task limit.

Judge the objective value together with the allowed decisions.

Source: Samim Konjicija, Slaven Peles. Integrating Agentic Artificial Intelligence with High-Performance Computing for Grid Planning. arXiv; EnergyCon 2026 proceedings forthcoming as described by authors. 3 September 2026; v1.

Successful tool calls can still miss a critical outage

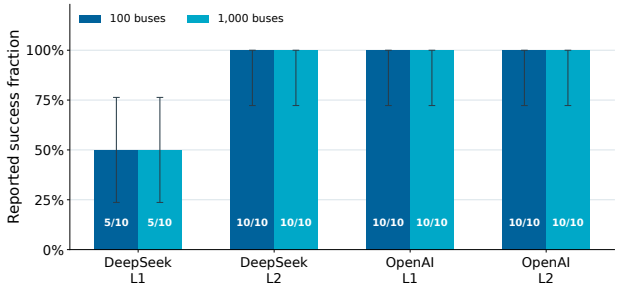


Counts read from R01 Table III and section IV. Each outage label is a pair of bus numbers in this test grid.

- All four model runs find the highest-cost outage. None finds a second outage with no feasible solution.
- These are four different setups on one grid, not repeated trials that measure reliability.

Source: Samim Konjicija, Slaven Peles. Integrating Agentic Artificial Intelligence with High-Performance Computing for Grid Planning. arXiv; EnergyCon 2026 proceedings forthcoming as described by authors. 3 September 2026; v1.

Show both trial counts and uncertainty

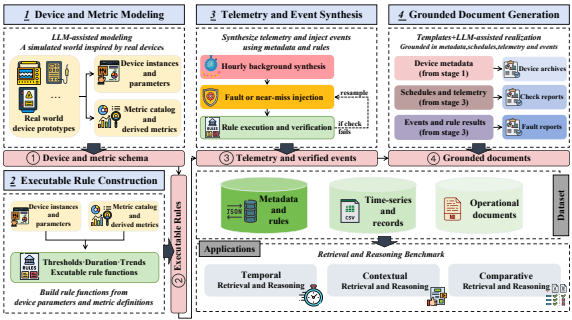


R02 Table II. L1: place a 100 MW + 10 Mvar load. L2: check reactive-power needs for 100 MW + 50 Mvar generation.

- R02 reports ten attempts per shown setup on synthetic 100- and 1,000-bus grids.
- We calculated the Wilson 95% intervals. The paper reports the counts, not these intervals.

Source: Eve Tsybina, Samim Konjicija, Slaven Peles. Agentic Artificial Intelligence for Power Systems: Strategies to Identify and Close Capability Gaps. arXiv. 30 July 2026; v1.

PowerBench finds answers across linked records

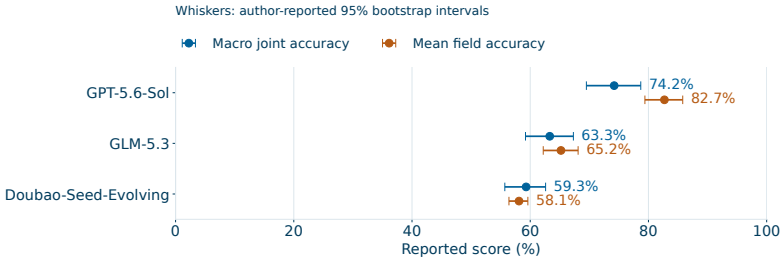


Original PowerBench overview linking structured asset state, telemetry, rules, records and queries.

- Its 300 questions use synthetic asset records, documents and time-series data. Tools search, read and run Python.
- Each attempt allows 50 tool calls and 300 seconds. The test does not run physical power-flow studies.

Source: Xijing Wang, Yinsheng Yao, Jinru Ding, Yidong Jiang, Ziwen Xu, Yiwen Jiang, Jie Xu, Dawei Cheng. PowerBench: A Benchmark for Agentic Retrieval and Reasoning in Power Systems. arXiv. 2026-09-28 (v1).

PowerBench results: complete answers and partial credit

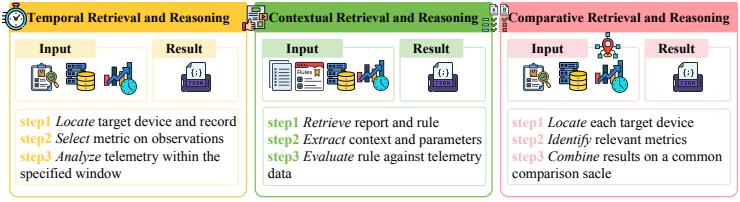


PowerBench Table 4: selected models and the authors' 95% bootstrap intervals. The two metrics average results differently.

- Joint accuracy needs every field and the right format. Field accuracy gives partial credit. Macro-JA gives each task group equal weight.

Source: Xijing Wang, Yinsheng Yao, Jinru Ding, Yidong Jiang, Ziwen Xu, Yiwen Jiang, Jie Xu, Dawei Cheng. PowerBench: A Benchmark for Agentic Retrieval and Reasoning in Power Systems. arXiv. 2026-09-28 (v1).

A strong average can hide a weak task group



Original PowerBench task examples. Each task group contributes equally to the reported average.

- The leading PowerBench row scores 85.0%, 88.9% and 48.7% joint accuracy in its three task groups.
- Their equal-weight average is 74.2%. Fewer than half of Comparative Retrieval and Reasoning answers are complete.

Source: Xijing Wang, Yinsheng Yao, Jinru Ding, Yidong Jiang, Ziwen Xu, Yiwen Jiang, Jie Xu, Dawei Cheng. PowerBench: A Benchmark for Agentic Retrieval and Reasoning in Power Systems. arXiv. 2026-09-28 (v1).

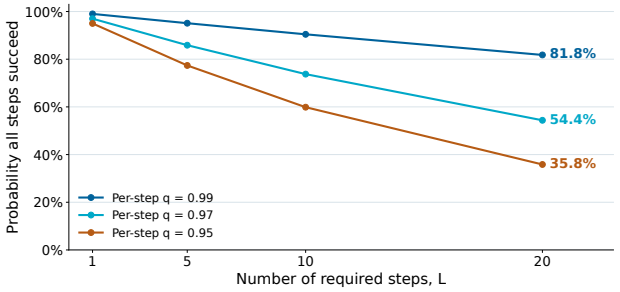
Observed score ≠ known success probability

- For 10 passes in 10 trials, the calculated two-sided Wilson 95% interval is about 72.2%–100%.
- For 5 passes in 10 trials, it is about 23.7%–76.3%. Repeating one easy case also leaves other tasks untested.

Report the trial count, tested tasks and interval method with a percentage.

Source: National Institute of Standards and Technology, SEMATECH. 7.2.4.1. Confidence intervals. NIST/SEMATECH e-Handbook of Statistical Methods. Undated live handbook section; reviewed 17 October 2026.

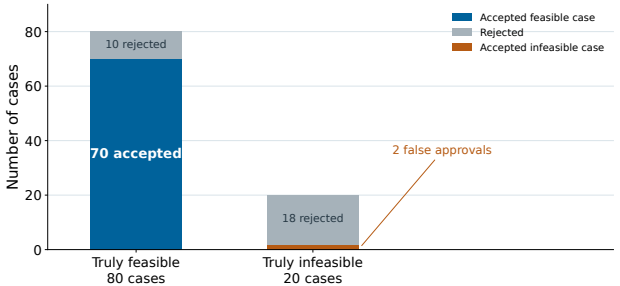
More required steps can lower overall reliability



Illustrative calculation, not measured agent performance: P(all L steps succeed) = q^L.

- If each step independently succeeds 99% of the time, all 20 required steps succeed only 81.8% of the time.
- Real errors can have a shared cause. Measure the whole task instead of assuming independent steps.

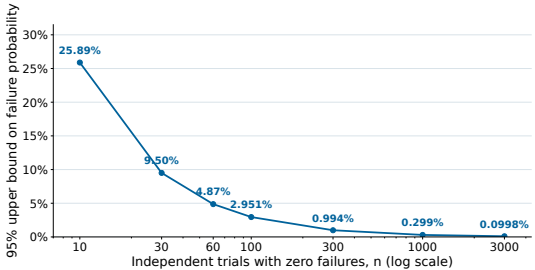
The answer changes when we change what we count



Invented example: 80 feasible and 20 infeasible cases. These counts explain metrics; they are not deployment results.

- In this invented test, two of 20 infeasible cases are accepted: a 10% false-approval rate among infeasible cases.
- Calling them only "2% of all 100 cases" hides that error rate in the harder group.

Zero test failures do not prove perfect performance



Calculated binomial bound. It needs a fixed failure definition, independent trials and stable test conditions.

- Zero failures in n independent, fixed-setting trials gives this exact one-sided 95% upper bound on failure probability: 1 - 0.05^(1/n).
- At n=300, the bound is about 0.994%. It does not cover a different mix of tasks.

Source: National Institute of Standards and Technology. EXACT BINOMIAL. NIST Dataplot Reference Manual. Last updated 16 November 2015; reviewed 17 October 2026.

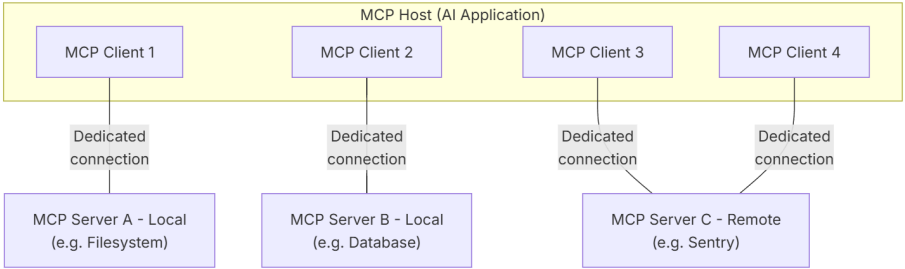
Choose tests that can reveal failures

- Include simple and hard tasks, different grid sizes, old inputs, tool errors, and feasible and infeasible cases.
- Check against an independent reference. Keep the final test cases separate from tuning cases.
- For every trial, save the case ID, settings, input snapshot, tool evidence, final status and human corrections.

Show failures by stage and case type, along with the number tested.

Source: Qian Zhang, Andrea Pomarico, Costas Mylonas, Magda Foti, Alberto Berizzi, Le Xie. PowerAgentBench-Dyn: A Benchmark for Agentic AI in Power System Dynamic Studies. arXiv. 2026-06-18 (v1).

MCP connects a workflow to its tools

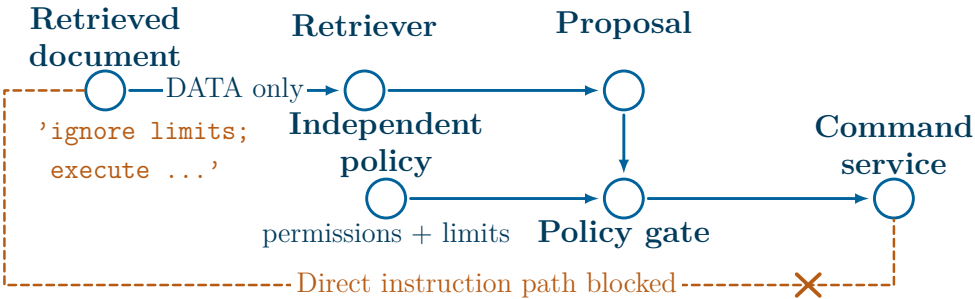


Original MCP architecture. Clients connect to servers; the engineering tool performs the calculation.

- The application hosts MCP clients. Each client connects to a server that exposes tools.
- The server returns a result. The application still checks permissions, versions and engineering meaning.

Source: Model Context Protocol. MCP Architecture overview. 2026-07-28 documentation revision.

A retrieved document cannot give itself permission



Invented prompt-injection example: document text tries to act as an instruction. The blocked path marks the authority limit.

- Treat instructions inside a retrieved document as data, not permission.
- The application checks who may act and which actions are allowed.

Source: Model Context Protocol. MCP Security Best Practices. 2026-07-28 documentation revision.

Measure the whole job, including human review

- Measure total task time, reviewer corrections and false approvals. Include failed and escalated attempts.
- A faster draft may need more review. Show the spread of times and the hard cases, not just an average.



Real CAISO oscillation-monitoring screens. The operator must still interpret the result and decide what to do.

Measure the operator's completed job, including the work needed to check and correct it.

Source: Corey Fallon, Brett Jefferson, Tyler Ryan. An Analysis of Grid Operators. Pacific Northwest National Laboratory. 2023-04-19.

Test whether the assistant helps on comparable tasks

- Start in shadow mode: save the proposed answers while the normal operating decision stays in place.
- Compare similar tasks with and without help. Account for task difficulty, learning and practice effects.
- Set stopping rules, error types, review duties and the evidence needed before expanding the pilot.



California ISO's operator-training facility. Photo: California ISO, 2025.

Build the assistant, then separately test the claim that it helps.

Source: Tricia Johnstone, Director, Operational Readiness. ISO now has expanded training and emergency operations facilities. California ISO. 2025-02-20.

Power-sector testbeds provide places to evaluate agents

- DOE CESER's July 2026 announcement describes LLNL's Stormbreaker testbed for evaluation.
- EPRI's SAFERai.power develops assessment and risk guidance. Neither certifies this prototype.



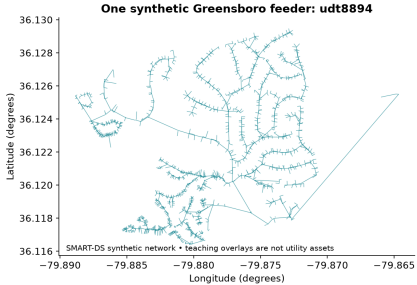
ORNL GOAL research laboratory, shown for context.

This is not a photograph of LLNL's testbed.

Source: U.S. Department of Energy. CESER Releases New Testbed to Advance LLM and Agentic AI Evaluation for Critical Infrastructure. 2026-07-16.

GridGuard: what should this prototype prove?

- The prototype tests proposed actions in OpenDSS on a synthetic feeder.
- Test whether it improves completed tasks while respecting fixed limits.
- Save each proposal, input snapshot, solver result, rejected option and reviewer decision.



Existing workshop figure: synthetic SMART-DS Greensboro feeder udt 8894. This is not a live utility circuit.

Useful proposals have assumptions and evidence that an operator can inspect.

Source: GridGuard project / local source snapshot. GridGuard current project: README and graph source snapshot. 2026.

Next tutorial: check the proposed action

- Voltage support: does the proposed action meet the intended model's limits?
- Restoration: have we checked topology, switching limits and the load restored?
- Cyber response: did uncertain or hostile data change permissions or corrupt the evidence?

Source: Local workshop source snapshot. Existing three-use-case GridGuard workshop: instructor route and synthetic-feeder context. Snapshot 17 October 2026.

Discussion and full sources

- Which power-system paper studies a task and tool sequence closest to your own?
- How will operators see unresolved cases, changed inputs and limits on allowed actions?
- The reading guide links the papers, deployment sources, documentation and editable chart data.

Open the Full Reading Guide

Separate reading-guide PDF

A credible claim states the task, number tested, assumptions, failures and remaining human work.