Agentic AI: Three Operator Cases

Understand the problem, follow the agent and review the evidence.

90 minutes / three what-if discussions

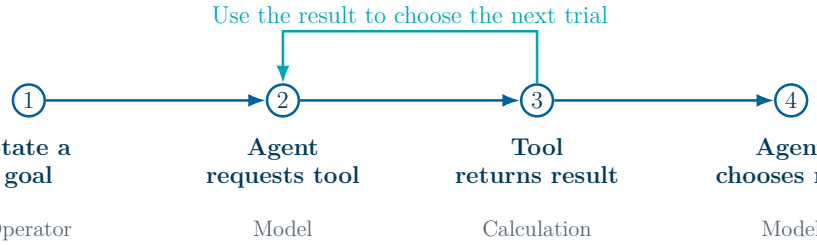
Alaa Selim
MIT/Dartmouth

Mohamed Chehada
UT Austin

IEEE ComSoc IEEE Communications Society | IEEE

Independent educational material / IEEE ComSoc theme adaptation

An agent chooses what to try next



Teaching diagram: the model requests a tool and uses the returned evidence.

- Give the agent a goal. It requests a tool, reads the result and chooses the next step.

Source: Monica Lam. CS224V Agentic AI, Lecture 1: Introduction. Stanford University, Fall 2026, p.2.

The agent asks; the tool calculates

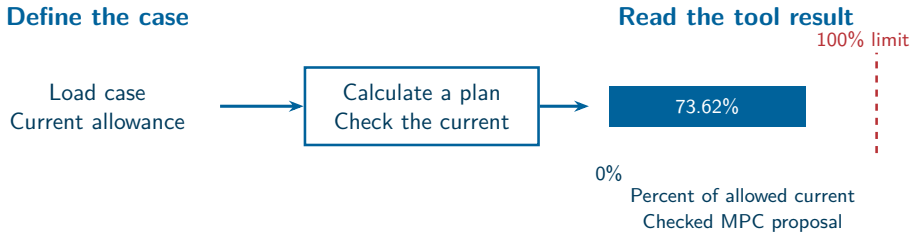


Original Stanford figure: one function call and a multi-step workflow.

- A tool does one job. An agent can connect several jobs using their results.

Source: Monica Lam. CS224V Agentic AI, Lecture 1: Introduction. Stanford University, Fall 2026, p.2.

Solar settings: which option should we try?



This current check passes. Review voltage and model coverage too.

The tool calculates; the agent coordinates the next request.

- Compare possible settings, then check voltage and line current.
- The agent can organize these trials and report the checked options.

Give the agent a goal and clear limits



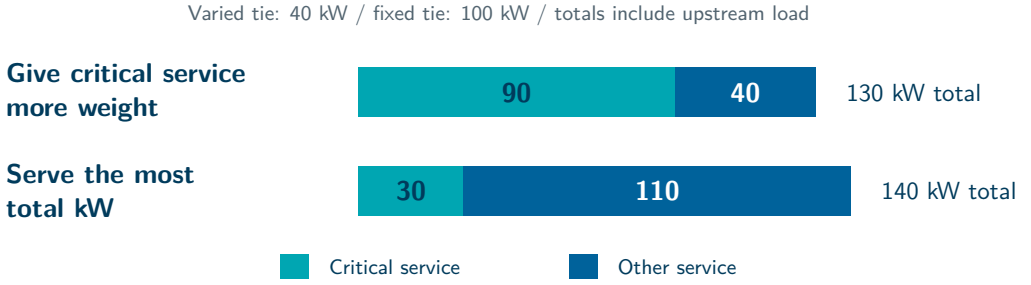
1	2	3
State the goal	**Name allowed tools**	**Set stop rules**
What should be achieved?	What may be requested?	When is review needed?

Grid teaching adaptation of Stanford's task-specification concept.

- State the goal, allowed tools and stopping rules.
- Keep the operator's review step explicit.

Source: Monica Lam. CS224V Agentic AI, Lecture 1: Introduction. Stanford University, Fall 2026, p.40.

Restoration: who should get power first?



Same network and limits, different service goals.

- The operator sets the service priority.
- That priority changes the study the agent requests.

Keep three jobs clear



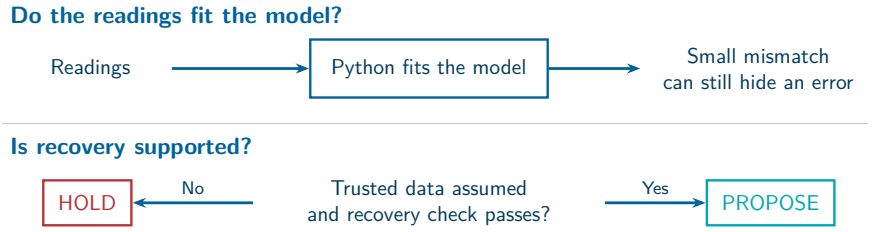
1	2	3
Agent requests	**Tool calculates**	**Operator decides**
Choose the study	Return the result	Review the next step

Adapted principle: language, calculation and human review have different jobs.

- The agent requests a study. The tool calculates. The operator reviews the result.

Source: Monica Lam. CS224V Agentic AI, Lecture 1: Introduction. Stanford University, Fall 2026, p.53.

Wrong readings can still fit a model

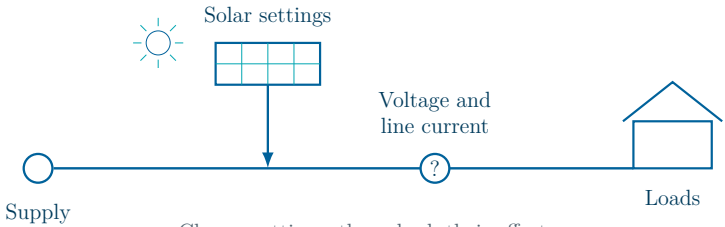


The extra readings are assumed trustworthy. The numerical tool does not authenticate them.

Model fit and trusted information answer different questions.

- A good fit alone does not prove that the readings are true.
- The agent can organize checks and flag missing evidence.

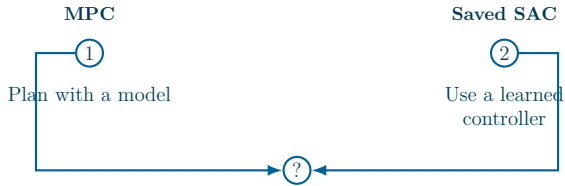
Case 1: which solar settings should we use?



Concept sketch: choose settings, then calculate their effects.

- Use solar generation while keeping voltage and line current within the allowed range.
- An agent can organize the trials and return checked options.

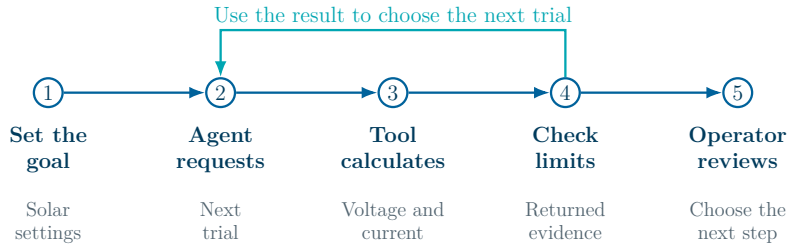
Two tools suggest different settings



MPC: model predictive control. SAC: soft actor-critic. Both proposals need the same checks.

- MPC plans with a model. SAC uses a previously trained controller.
- Treat both outputs as proposals that need the same electrical checks.

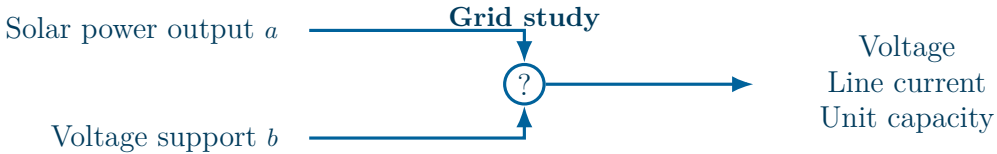
How the agent helps with solar settings



The agent uses returned checks to choose the next permitted step.

- The agent requests a study and reads the returned checks.
- It can try another allowed option or ask the operator for help.

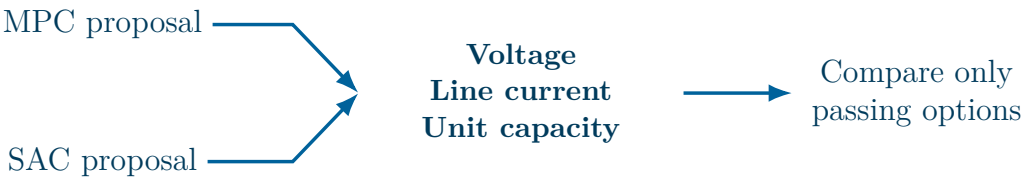
Two solar settings change the result



Active power and voltage support share equipment capacity.

- One setting controls solar power output. The other provides voltage support.
- The tool checks voltage, line current and the solar unit's capacity.

Both proposals face the same checks

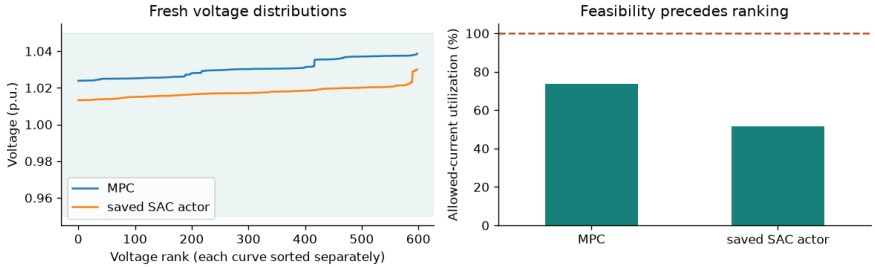


The agent requests; the tool calculates and checks

Both proposals need the same voltage, current and capacity checks.

- The tool calculates the effect of each proposal.
- Compare their scores only after voltage, current and capacity checks pass.

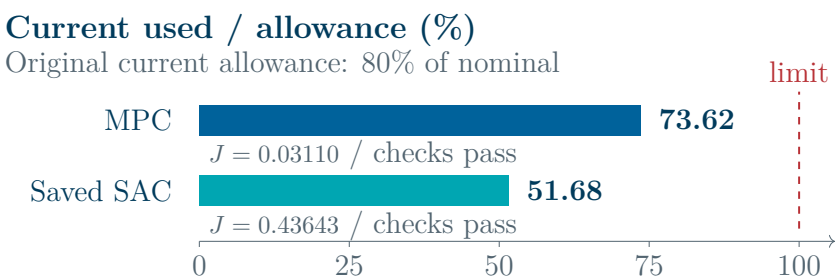
Read the voltage and current checks



Calculated example at the original current allowance.

- Left: voltages should stay in the allowed band.
- Right: current should stay below 100% of the allowance.

Both pass, but the trade-off is different



Calculated example: both electrical checks pass.

- MPC uses 73.62% of the current allowance; SAC uses 51.68%.
- MPC has the lower overall score because current is only one part of the goal.

What if we allow less current?



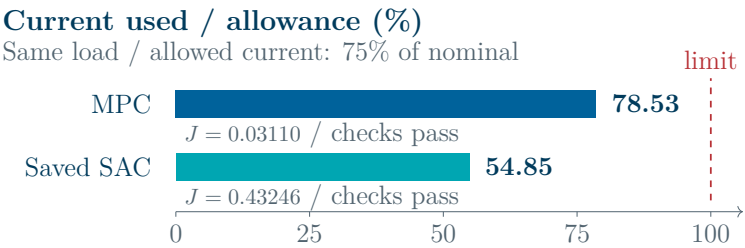
Allowed current as a share of nominal current

What needs to be checked again?

Discuss the scenario, then advance to the calculated result.

- Reduce the allowance from 80% to 75% of nominal current.
- Which proposal might fail? Which would you choose if both pass?
- What must the agent ask the tool to check again?

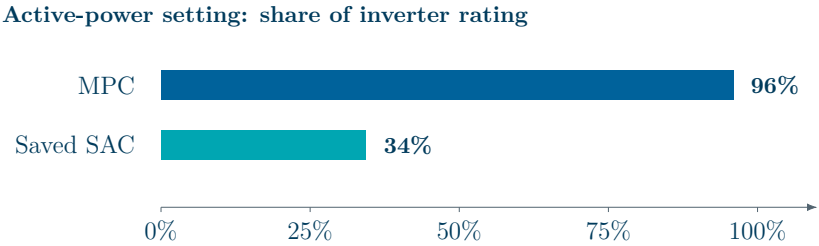
Both proposals pass the new current check



Same load; the allowed current has been reduced to 75% of nominal.

- MPC has the lower score. SAC uses less current and less solar output.
- These results cover the medium-voltage model. A wider grid needs a wider study.

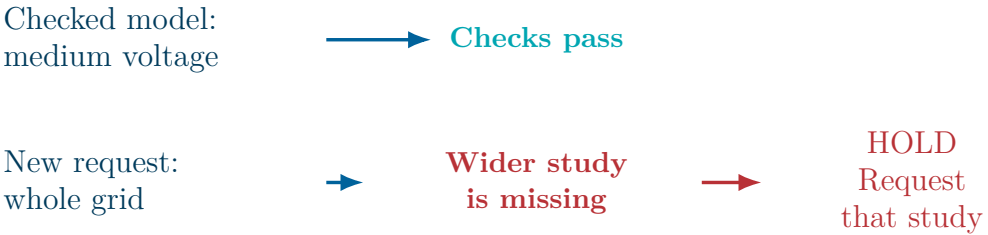
Less current does not mean the best option



Active-power settings use inverter rating as their base.

- Here SAC also uses less solar output.
- The operator's goal values solar generation as well as operating limits.

A PASS covers only the checked model



A passing result applies to the model and conditions actually checked.

- A study of the medium-voltage network cannot confirm a whole-grid claim.
- The agent should request the missing study or return HOLD.

Recorded example: the first option fails

RECORDED REPLAY · llama3.1:8b · voltage · call 1

Input facts	Forecast load multipliers: [1.0, 0.94, 1.06]. Current allowance: 80% of nominal; voltage limits: [0.95, 1.05] p.u.
Llama choice	`VERIFY_SAC_FIRST`
Exact public reason	“A numerical SAC proposal is available; its fresh first-action check has not yet been requested.”
Recorded result	OpenDSS: 111.966607% of the allowed current; 2 current violations. Deterministic check: REJECT; failure = thermal.

Recorded full-feeder study; different inputs from the earlier comparison.

- Llama requests a check of the saved SAC proposal.
- The numerical tool reports too much current and rejects it.

Recorded example: another option is checked

RECORDED REPLAY · llama3.1:8b · voltage · call 2

Input facts	The supplied failed checks are: thermal. An MPC plan already exists; its recovery-eligibility failure list is empty.
Llama choice	`USE_MPC_PLAN`
Exact public reason	“The MPC plan has been generated and is eligible for a numerical check. This is the next allowed step in the recovery process.”
Recorded result	OpenDSS: 99.604410% of allowed current; no recorded violations. Proposed P = 2387.157 kW; Q = +325.738 kvar.

Recorded full-feeder study: the requested fallback passes the checks.

- Llama requests a check of the existing MPC fallback.
- The returned electrical checks pass for this recorded case.

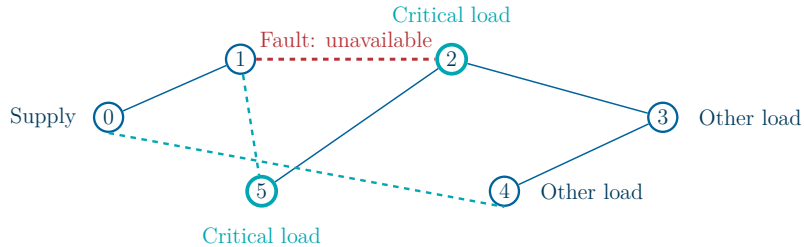
Recorded example: the operator receives a proposal

RECORDED REPLAY · llama3.1:8b · voltage · call 3	
Input facts	The full-feeder first-action check passed; voltage is 1.001783–1.035026 p.u. Protection, transients and implementation authority remain outside this model.
Llama choice	`REVIEW_EVIDENCE`
Exact public reason	“The model has converged and the numerical evidence indicates that the control actions are within the allowed limits. Reviewing the evidence will provide more insight into the model's performance and the selected control actions.”
Recorded result	Numerical status: MODEL_RECOMMEND; recorded human acknowledgment: PENDING. Acknowledged = False; dispatch authorized = False; equipment operated = False.

Recorded workflow result; human review remains pending.

- The agent returns the checked proposal and its conditions.
- The operator still reviews it before any authorized action.

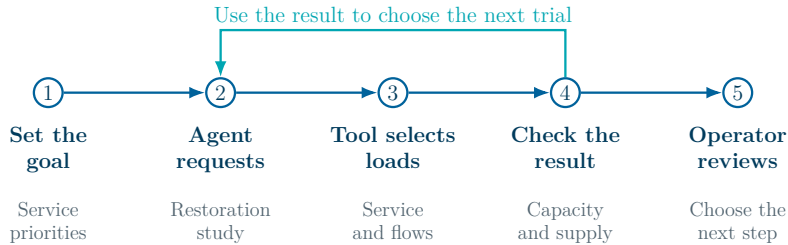
Case 2: who gets power after a fault?



Six-node example. Cyan marks critical loads; dashed links are backup paths.

- A fault removes a supply path. Available capacity limits what can be restored.
- The operator sets priorities; the agent can request and check restoration options.

How the agent helps restore service



The operator sets the priority. The agent requests the matching study.

- Choose the goal: critical loads first, or the most total power.
- The agent can request a study, read the trade-off and ask for a new study when limits change.

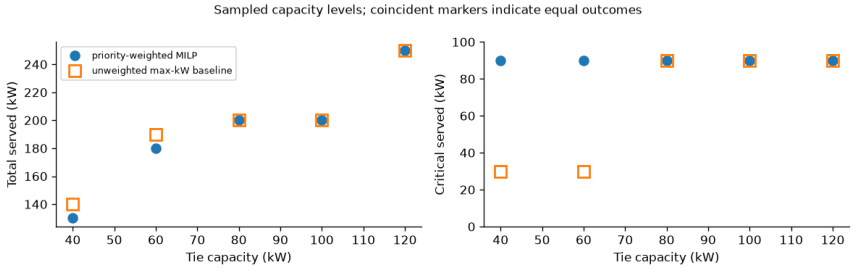
The service goal changes the study request

Operator's goal	Agent requests	Tool calculates
Protect critical service	Critical-first study	130 kW total 90 kW critical
Restore the most power	Most-total-power study	140 kW total 30 kW critical

Two possible operator goals for the same fault and available network.

- Critical-first gives more weight to essential service.
- Most-total-power aims to restore the largest amount of power.

The goal changes who receives power



Calculated restoration results for the same network and limits.

- Critical-first: 130 kW total, including 90 kW for critical loads.
- Most-total-power: 140 kW total, including 30 kW for critical loads.

Check the restoration plan



1	2	3	4
Fault stays isolated	**Supply paths connect**	**Capacity checks pass**	**Report service**
Unavailable branch	Restored loads	Within model limits	Total and critical kW

Active-power model checks. Voltage, protection and field switching require separate work.

- Keep the faulted branch out of service. Check supply paths and capacity.
- Voltage, protection and field switching need separate studies.

What if the backup path can carry more power?

40 kW → **180 kW**

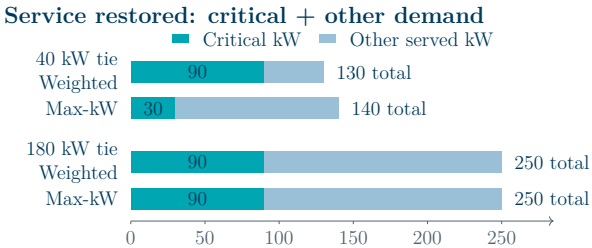
Backup path capacity / same fault and loads

Will the two service goals still disagree?

Discuss the scenario, then advance to the calculated result.

- Increase its capacity from 40 kW to 180 kW. Keep the fault and loads the same.
- Would the two service goals still give different results?
- What should the agent recalculate before reporting the new plan?

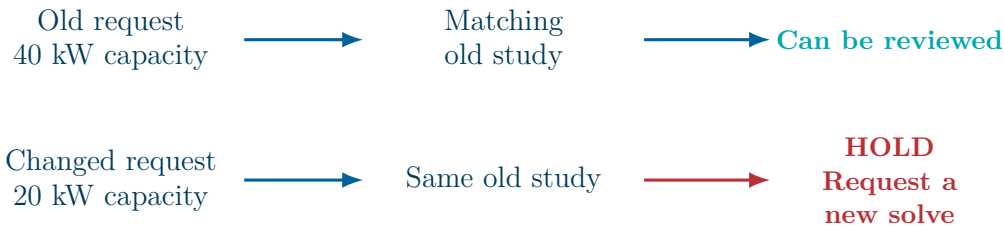
More capacity removes this service trade-off



Calculated results with the same fault, loads and service goals.

- At 40 kW, critical-first serves 90 kW of critical load instead of 30 kW.
- At 180 kW, both goals serve 250 kW total, including all 90 kW of critical load.

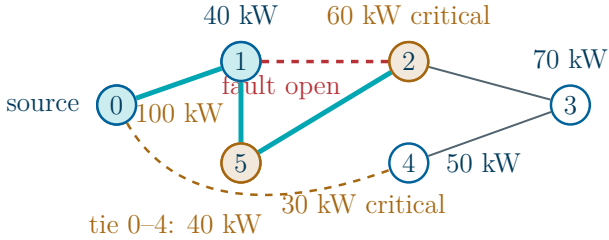
A new limit needs a new study



A changed limit requires evidence calculated for the new request.

- Now reduce the allowed backup capacity from 40 kW to 20 kW.
- The old result no longer supports the request. The agent must request a new study.

The agent keeps the operator's goal



Picked up:
1, 2, 5
130 kW total

Critical service
90 kW

The operator's priority stays fixed while the tool selects a feasible plan.

- The operator sets priorities. The solver selects loads.
- The agent can request the solve and explain the returned service trade-off.

Recorded example: another plan restores more service

RECORDED REPLAY · llama3.1:8b · restoration · call 14	
Input facts	E01 failed its thermal check; E02 passed. E02 serves 1242.777 kW with 65.938897% tie loading.
Llama choice	`SHED_075`
Exact public reason	“The candidate E01 has a thermal failure, and the lowest cost passed candidate E02 has a cost of -5509.135240000002. Selecting SHED_075 will test 75% disconnection of the ordered noncritical load objects, which may improve the cost.”
Recorded result	SHED_075 produces E03: 1843.614 kW served and 474.040 critical kW. E03 passes at 91.867215% tie loading; cost = -6109.971300 (smaller is better).

Recorded larger-feeder example, separate from the six-node network.

- In this separate larger-feeder study, Llama requests another plan.
- The tool reports more service and passing electrical checks.

Recorded example: the tool rejects a trial

RECORDED REPLAY · llama3.1:8b · restoration · call 15	
Input facts	E03 is the lowest-cost passed candidate; one refinement evaluation remains. SHED_065 means disconnecting 65% of ordered noncritical load objects, not 65% of kW.
Llama choice	`SHED_065`
Exact public reason	“The candidate E01 has a thermal failure, and the lowest cost passed candidate E03 has a cost of -6109.9713, which is lower than the objective's target. Selecting SHED_065 will test 65% disconnection of the ordered noncritical load objects, which may improve the cost.”
Recorded result	E04 reaches 111.009413% tie loading: REJECT, thermal failure. The rejected trial has no eligible score; its larger served load cannot make it acceptable.

Recorded larger-feeder example: the overloaded trial is rejected.

- The next trial carries too much current.
- It cannot replace the previous passing option.

Recorded example: a wrong ranking is corrected

RECORDED REPLAY · llama3.1:8b · restoration · call 16

Input facts	E02 cost = -5509.135240; E03 cost = -6109.971300; smaller is better. Only the passed candidates E03 and E02 are offered for ranking.
Llama choice	`E02, E03`
Exact public reason	“The candidates were ranked based on their scores, with the lowest score (best) first. E02 has a score of -5509.135240000002, and E03 has a score of -6109.971300000003.”
Recorded result	Returned language order: E02 → E03. Correct numerical order: E03 → E02. The guard retains E03 and rechecks it as E05: 1843.614 kW served; no recorded failures.

Recorded larger-feeder example. Lower cost is preferred, including negative costs.

- Llama puts two passing plans in the wrong cost order.
- Python keeps the lower-cost plan and checks it again.

Case 3: can we trust these readings?

Some readings may be wrong

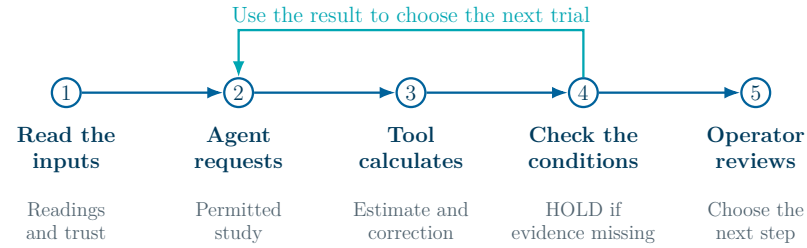


A good fit is not the same as true data

Concept sketch using a small four-bus model.

- Wrong readings can still look consistent with the model.
- The agent can request checks, but action still needs a trusted basis.

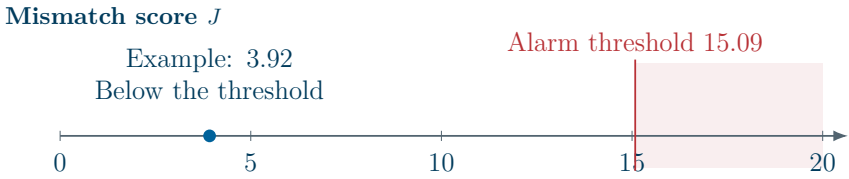
How the agent helps with suspicious readings



The agent requests checks and preserves missing evidence.

- The agent can request the checks and keep missing evidence visible.
- The operator reviews trust, model coverage and the proposed correction.

Mismatch is only one check



A low mismatch does not authenticate a reading

Mismatch measures model fit; state error measures estimation accuracy.

- Mismatch J measures how well the model fits the readings.
- The simulation lets us measure how wrong the estimate is (state error).

Two patterns of wrong readings

One reading changes



May conflict with
the remaining readings

Several change together



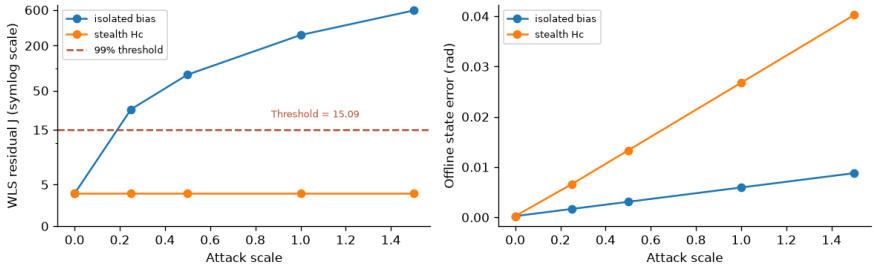
May still fit
the grid model

Model consistency and data truth are different checks

Concept illustration of two patterns used in the numerical comparison.

- One wrong reading may conflict with the other readings.
- Coordinated changes may still look consistent with the model.

Read the alarm and the estimation error



Calculated model-fit and state-error results.

- Left: mismatch and its alarm threshold. Right: error against the simulation's known truth.
- A quiet alarm can appear alongside a growing error.

A coordinated change can escape the alarm

One biased reading

$J = 277.50$

↓

Alarm

Coordinated change

$J = 3.92$

↓

No alarm

State error: 0.02678 rad

The estimate is still wrong

The known state is available inside the simulation for this comparison.

- At scale 1, one biased reading gives mismatch J=277.50 and an alarm.
- A coordinated change gives J=3.92, no alarm, and 0.02678 rad of state error.

What if the data change is small?



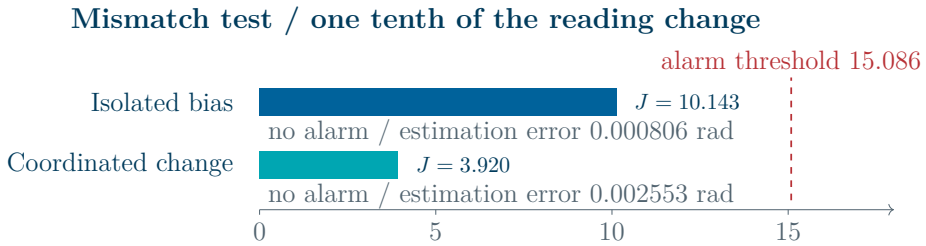
Size of the reading change

What can a quiet alarm tell us?

Discuss the scenario, then advance to the calculated result.

- Reduce the change to one tenth of the reference size.
- Will the alarm always detect the error?
- What else would the operator need before trusting the readings?

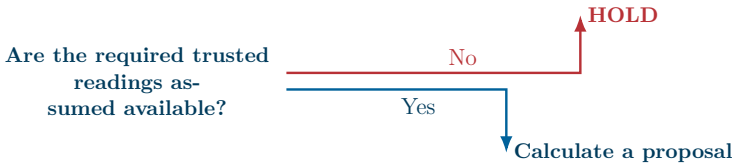
A quiet alarm can still hide an error



Calculated results at one tenth of the reference change size.

- Neither example triggers this alarm at one tenth of the reference size.
- Both estimates still contain error. No alarm alone does not establish data trust.

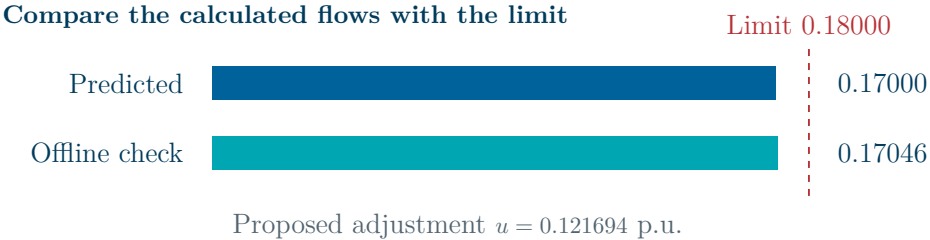
Trusted information is needed before recovery



Additional trustworthy readings are assumed in this example.

- Without the required trusted readings, the response is HOLD.
- With readings assumed trustworthy, the tool can calculate a proposal for review.

The tool calculates a small correction



Per-unit values share the same model base.

- The solver finds the smallest allowed adjustment that meets the checked branch limits.
- Predicted flow: 0.17000 p.u. Offline check: 0.17046 p.u. Limit: 0.18000 p.u.

Recorded example: request the recovery tool

RECORDED REPLAY · llama3.1:8b · cyber · call 7

Input facts	The structured task already authorizes WLS_TRUSTED_REDISPATCH. Trusted sensors are explicitly assumed; no live utility measurements are supplied.
Llama choice	`WLS_TRUSTED_REDISPATCH`
Exact public reason	“This tool is authorized for the given use case and scope, and it is designed for a specific DC redispatch example with trusted measurements.”
Recorded result	The numerical tool proposes u = 0.121693667 p.u.; predicted flow12 = 0.170000 p.u. Offline synthetic-truth flow12 = 0.170461 p.u.; rating = 0.180000 p.u.

Recorded recovery request and numerical tool result.

- Llama requests the supplied recovery tool.
- Python calculates a correction using the assumed trusted readings.

Recorded example: keep human review

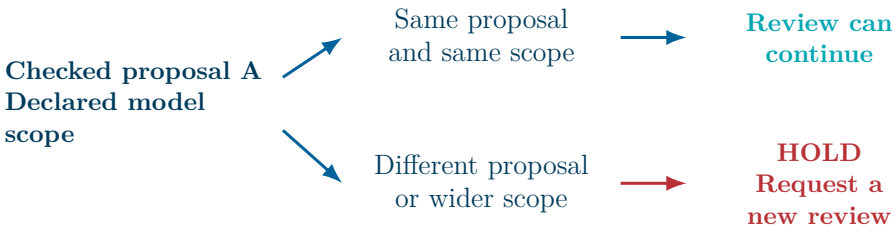
RECORDED REPLAY · llama3.1:8b · cyber · call 8

Input facts	The model receives the actual numerical summary and passed_modeled_checks = True. No communication authentication, AC feeder or operating-permission check is included.
Llama choice	`REVIEW`
Exact public reason	“The provided JSON output indicates that the numerical workflow has completed successfully, and the results are within the expected scope. The 'passed_modeled_checks' field is set to True, indicating that the modeled checks have been passed. Therefore, the choice is to REVIEW the results to ensure they are accurate and within the expected range.”
Recorded result	Numerical verification: PASS. Language choice: REVIEW. Overall result: REVIEW. Human acknowledgment: PENDING; dispatch authorized = False; equipment operated = False.

Recorded workflow result; human review remains pending.

- The model checks pass; human review is still pending.
- Data trust and operating permission still need review.

Review must refer to the same proposal

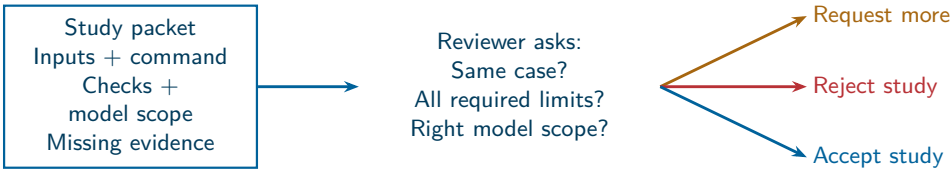


Reviewing a study does not operate equipment

Illustrative proposal IDs. Review is separate from equipment operation.

- The proposal ID and model scope must match the checked study.
- A different proposal or wider scope needs a new review.

The operator chooses the next step



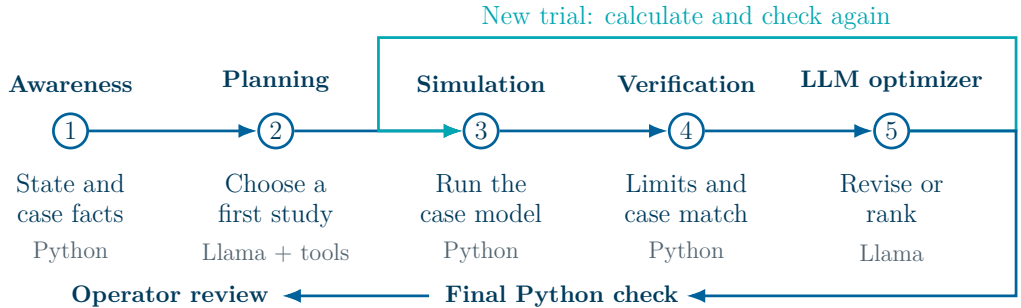
Accepted study $\neq$ permission to operate equipment.

Proposed advisory pilot — teaching design

The operator reviews the study and chooses the next authorized step.

- Review the case, tool results, checked limits and missing work.
- Request another study, reject the proposal or take the next authorized step.

Multi-agent framework used in these cases



Implemented five-role workflow with a case-specific numerical tool.

- The five roles share the same case facts and checked results.
- One Llama model handles planning and refinement. Python calculates and checks.

Explain one case in your own words

Solar settings	**Restoring service**	**Suspicious readings**
Which settings meet the goal?	Who gets power under the limits?	What evidence can be trusted?

Agent requests a permitted study → **Tool returns numerical evidence** → **Operator reviews the next step**

Explain the idea first. Practice material is available separately.

- What is the operator's problem, and what should the agent request next?
- Which tool result supports the proposal, and what must the operator still decide?