AGNTCON + MCPCON JAPAN 2026

Carbon-Aware Agentic Engineering: Reducing Unnecessary Computation in AI Agent Workflows

Kouki Hama | NTT, Inc.

10 SEP 2026 | 11:15 JST | HALL C

ABOUT THE SPEAKER

I connect environmental impact with practical software engineering.

Kouki Hama

Senior Research Engineer, NTT, Inc.
Computer & Data Science Laboratories

FOCUS

Reduce software energy and carbon
Make software components traceable

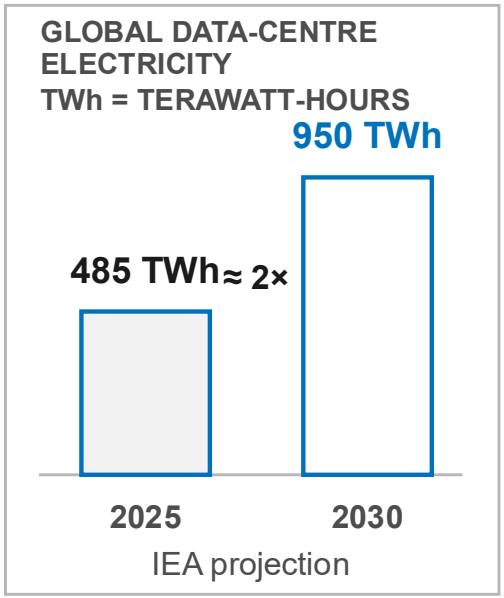
ENERGY + CARBON	SOFTWARE TRACEABILITY

Communities: OpenChain Japan
Eclipse SW360 co-lead

TABLE OF CONTENTS

Why AI Agent efficiency matters

Growing AI use makes avoidable Agent work a design concern.



WHY MULTI-STEP AI MATTERS

One visible request can trigger many Model calls, Tool calls, Retries, and larger Context. Avoidable steps multiply the work.

The exact energy depends on the task, model, infrastructure, and grid.

REFERENCE USED THROUGHOUT THIS TALK

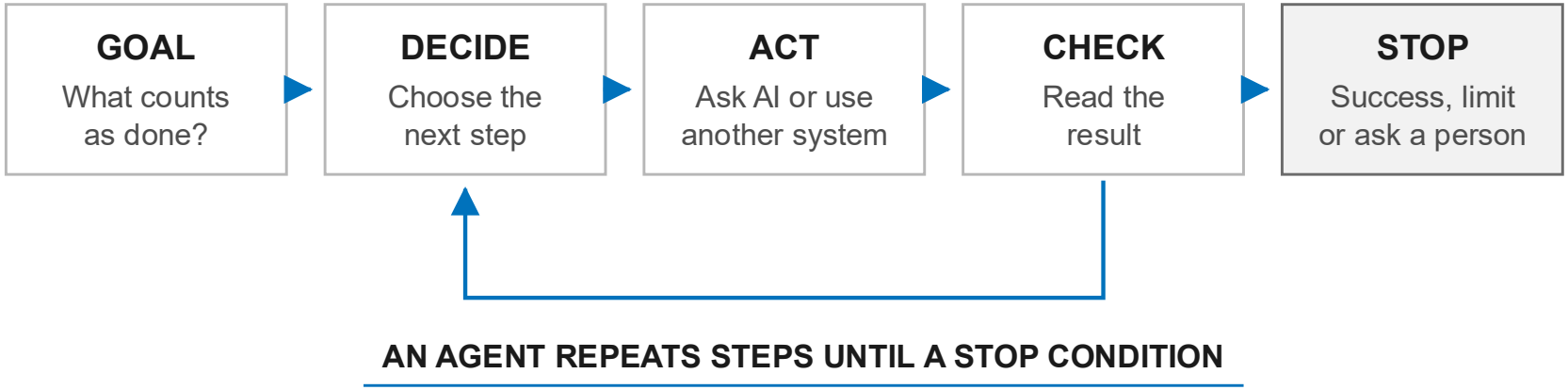
Lean and Green Agentic AI

An independent, vendor-neutral white paper by Navveen Balani. It maps waste and lean practices across six lifecycle stages and discusses measurement with SCI and SCI for AI.

WHITE PAPER = FRAMEWORK
GITHUB REPOSITORY = IMPLEMENTATION EXAMPLES

Use the white paper as the design frame: remove avoidable work first, then measure electricity and CO2.

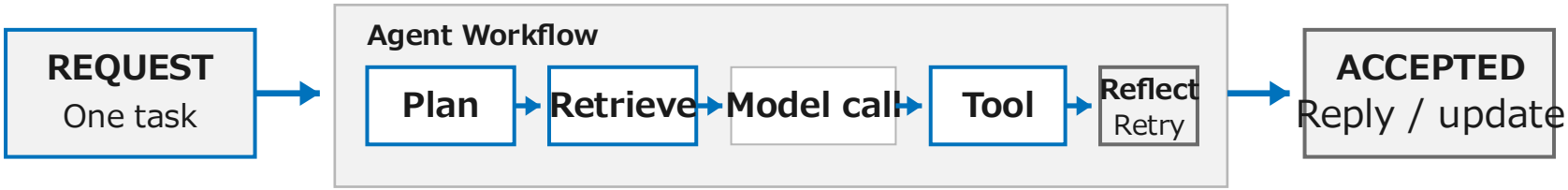
An AI agent repeats decisions and actions until it stops.



Use an agent only when a task really needs repeated decisions or actions.

WHY IT MATTERS

One request can trigger many operations inside an AI Agent



More operations mean more Model work, data, and runtime

Model call
Requests sent to the Model

Input / output data
Text sent and generated

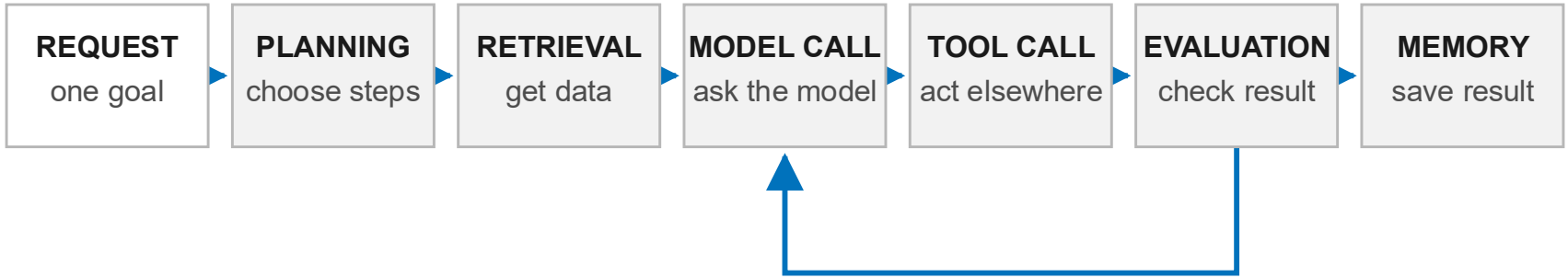
Runtime
Time compute resources stay active

Separate required operations from avoidable ones

Kim et al., HPCA 2026 | Poddar et al., NAACL 2025

Use Trace data to separate required operations from avoidable ones before measuring Energy.

One user request can trigger many separate operations.



EACH OPERATION CAN ADD COMPUTING, NETWORK TRAFFIC, TIME, AND COST

Environmental impact depends on all steps actually run, not the number of visible requests.

Source: Lean and Green Agentic AI (PDF)

How an AI Agent carries out a task

COMPARE TWO WAYS TO DO ONE TASK

The same AI-assisted task can use many more steps than necessary.

TASK: CREATE A 5-ITEM SUMMARY FROM 3 APPROVED DOCUMENTS

MULTI-AGENT + HANDOFFS

01 3 handoffs: one Agent passes work to the next

02 Each Agent rereads all documents

03 Largest model used at every step

04 Unbounded retry after failure

LEANER WORKFLOW

01 Code loads and validates the files

02 A small model extracts five fields

03 Human review only for uncertain items

SAME REQUIRED RESULT • SAME FIVE FIELDS • DIFFERENT NUMBER OF STEPS

Compare two paths only after both must produce the same required result.

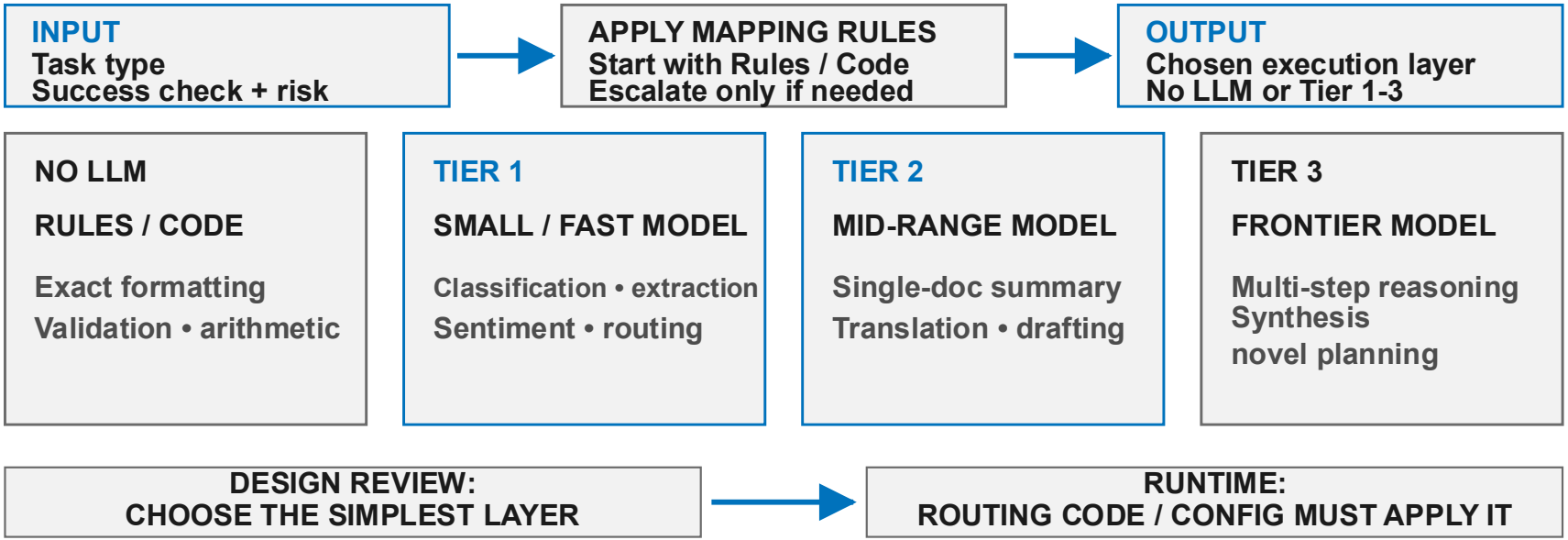
REVIEW THE AGENT'S WORK

Four questions show which agent steps can be simplified.

01	**What result counts as success?** Record: the success check	DEFINE SUCCESS
02	**Does this step need generative AI?** Could ordinary code or rules do it?	TRY CODE / RULES
03	**What is the simplest option that can still succeed?** Record: chosen model and needed information	USE SIMPLEST
04	**When must the agent stop?** Record: attempt limit and stop reason	LIMIT + STOP

Define the result, ask whether AI is needed, choose the simplest method, and decide when the agent stops.

Tier choice uses the task, success criteria, and risk - not an automatic switch.



A design review proposes the simplest layer; routing code or configuration applies it.

Source: lean-agentic-ai / right-size-your-model.md

Context is every input sent to a model; keep only what each step needs.

NTT

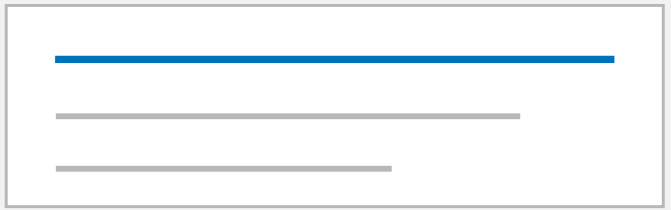
UNBOUNDED CONTEXT

Send full documents and history

more text to process • more noise

MINIMAL CONTEXT

Send only the relevant excerpt



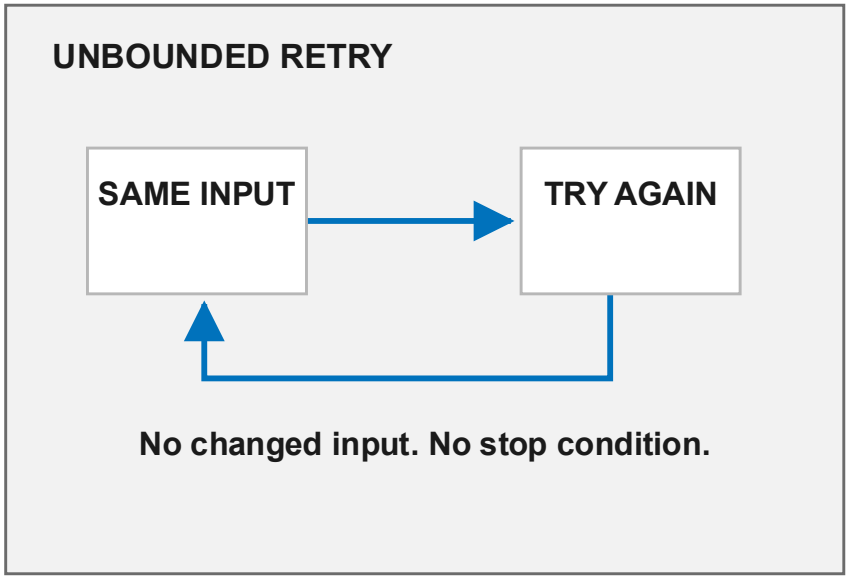
less text to process • clearer signal

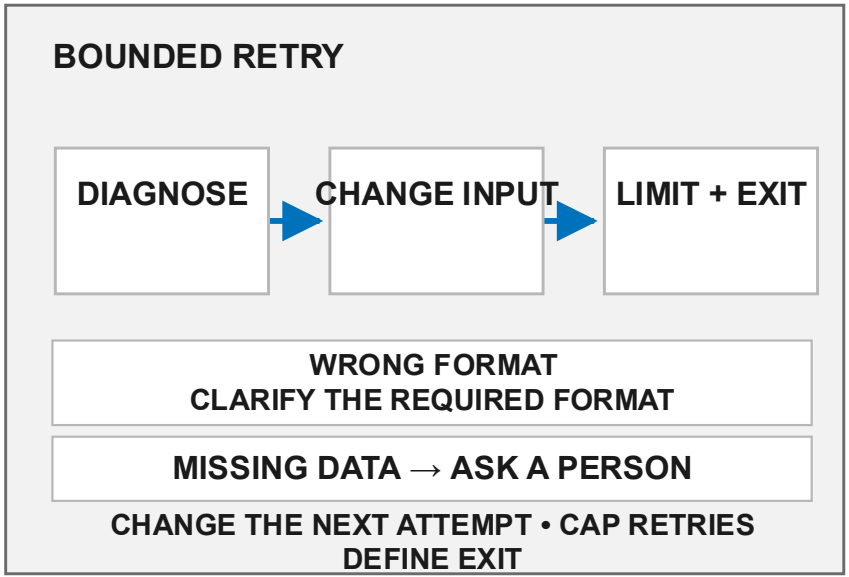
CONTEXT CHECK: NEEDED? • CAN IT BE SHORTENED? • DELETE AFTER USE?

Less input means less text to process and a clearer signal for the model.

Source: lean-agentic-ai / context-is-a-liability.md

A bounded retry changes the next attempt and defines a stop condition.



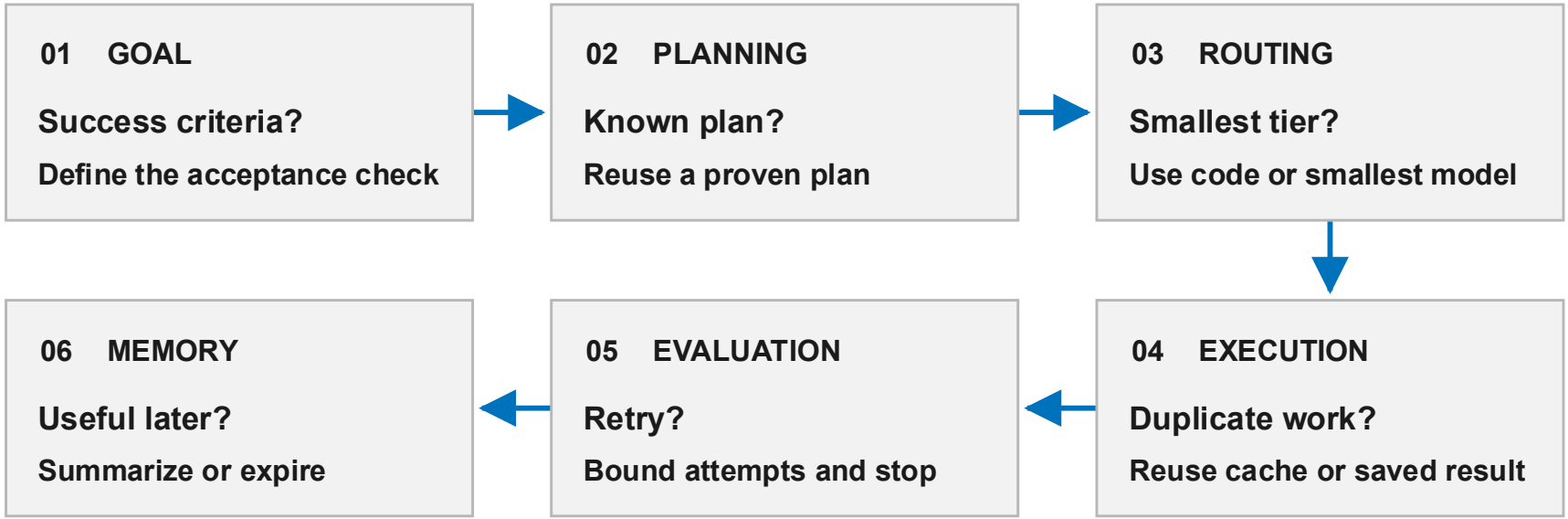


Correct the input or format, ask a person, or stop at the attempt limit.

Source: lean-agentic-ai / bounded-retries-and-more.md

Reduce and measure avoidable work

Review the full AI Agent lifecycle to find where avoidable work enters.



WHY IT MATTERS

The white paper organizes the review into Goal, Planning, Routing, Execution, Evaluation, and Memory.

Source: Lean and Green Agentic AI (PDF)

EXPERIMENT QUESTION

We tested whether Workflow design could reduce avoidable processing

Can a Lean Workflow pass the same checks with less work?

SIX DIFFERENT TASKS

1 Hours reply
2 Ticket route
3 Contact edit
4 Return rule
5 Summary
6 SLA escalation

Six tasks — not six repeated runs

TWO WORKFLOWS

Always-on
Retrieve → Plan → Answer → Reflect

Lean
Check conditions first; run only needed steps

Same input • pass criteria • deterministic Model stand-in

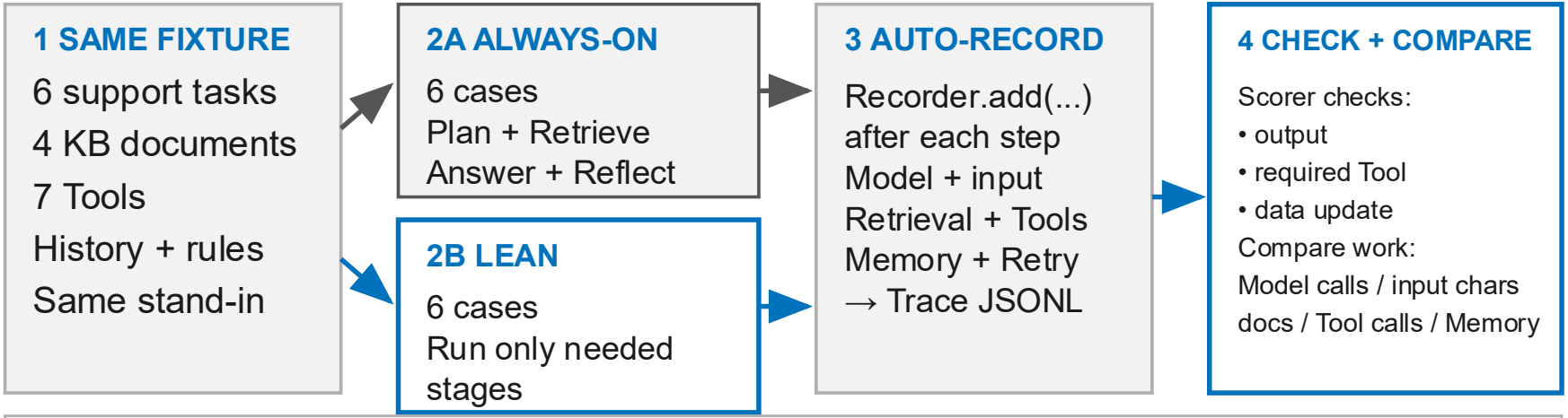
stand-in: fixed substitute; the same input always returns the same output

Compares Workflow paths — not real-LLM quality, Energy use, or CO2 emissions

These are six different tasks; both Workflows use the same inputs and acceptance criteria.

We ran the same six requests through two Workflows and logged every step.

6 tasks × 2 Workflow policies = 12 case executions



Controlled Trace experiment | External LLM API: 0 | Energy and CO_2: not measured

Only Workflow control changes; compare workload after both conditions pass.

The six requests cover different kinds of agent work.

NTT

01 BUSINESS HOURS
INPUT When is support open?
WORK Code gate; no Model call
PASS Mon–Fri, 09:00–17:00 JST

02 TICKET ROUTING
INPUT Route T-1007 to Billing
WORK Classification + route_ticket
PASS BILLING + required Tool ran

03 CONTACT UPDATE
INPUT Update Alice / email
WORK Extraction + structured update
PASS Profile state changed

04 RETURN POLICY
INPUT Return window for A100?
WORK Selective Retrieval
PASS Answer = 30 days

05 SESSION SUMMARY
INPUT Summarize the session
WORK Bounded Memory
PASS All 7 required facts retained

06 SLA ESCALATION
INPUT Should T-1007 escalate?
WORK Plan + 3 Tools + Reflection
PASS Enterprise + breach + impact

The workload covers code gates, Tools, Retrieval, Memory, and complex decisions.

For each case, the program recorded the path and checked the result.

INPUT **Route ticket T-1007 to Billing**

ALWAYS-ON

Full Memory → 4 documents → Profile Tool
Plan → Answer → Reflect → route_ticket

Model calls 3 | Retrieved docs 4
Tool calls 2 | Input 9,976 chars
PASS

LEAN

Routing Model call → route_ticket
Only the required context and Tool schema

Model calls 1 | Retrieved docs 0
Tool calls 1 | Input 836 chars
PASS

Recorder.add(...) → traces.jsonl **Scorer checks result + route_ticket**

Trace measures the work performed; Scorer confirms that the required business result was preserved.

RESULT

The Lean Workflow preserved outcomes with less internal processing

Acceptance checks: Always-on 6/6 passed | Lean 6/6 passed

Model-interface calls	Model-input characters	Retrieved documents
19 → 7	70,066 → 8,545	24 → 2
−63.2%	−87.8%	Only when required

Why 19 calls: 6 tasks × (Plan + Answer + Reflection) + 1 Retry

Reflection 6 → 1	Memory 4,160 → 262 chars	Workload ↓ ≠ Energy / CO_2 ↓

Same fixture and pass criteria; stand-in fixed; real-LLM quality not assessed.

We confirmed lower logged workload — not a measured Energy or CO2 reduction.

Next, measure actual Energy use and CO_2 emissions.

FUNCTIONAL UNIT one accepted task

1 TRACE	2 REAL MODEL	3 ENERGY	4 CO_2
Calls **Input size** **Path**	Provider tokens Latency / cache Model metadata	GPU / node Joules or Wh per accepted task	Energy × grid carbon intensity at runtime
CONFIRMED	NOT YET MEASURED		

Trace finds candidates; telemetry and Energy measurements verify the Green impact.

Measure Energy per accepted task; convert it using grid carbon intensity at runtime.

Summary: the abstract's three questions now have concrete answers.

01

Why can an AI Agent use more resources than one AI answer?

Planning, retrieval, model calls, tool calls, retries, and memory all add work.

02

How does removing unnecessary work support Green Software?

It prevents avoidable computation; then measurement shows what the remaining work costs.

03

What should an engineer check?

Need, execution layer, Context, Retry change, stop condition, and a fair before/after result.

Trace finds avoidable work; direct measurement must verify the Energy and CO2 impact.

Apply the method: map the work, remove waste, then verify the change fairly.

01 Is the success check defined?	02 Can we observe every model and tool call?
03 Can rules or code replace an LLM call?	04 Is Context limited to what this call needs?
05 Does each Retry change input and have a stop?	06 Do before and after pass the same success check?

KEEP AS PROOF: STEP DIAGRAM • ONE RUN RECORD • QUALITY + ACTIVITY COMPARISON

Keep evidence: the execution path, Trace log, and a before/after result that passes the same success check.

Sources cited in this presentation.

01 **SESSION ABSTRACT**
Carbon-Aware Agentic Engineering
sched.co/2QlEh
Session scope and promised questions

02 **WHITE PAPER + REPOSITORY**
Lean and Green Agentic AI
github.com/navveenb/lean-agentic-ai
Lifecycle and lean / green design principles

03 **ENERGY CONTEXT | IEA**
Key Questions on Energy and AI
iea.org/reports/key-questions-on-energy-and-ai
Electricity demand and AI workload context

04 **AGENT ENERGY | HPCA 2026**
Kim et al., The Cost of Dynamic Reasoning
doi.org/10.1109/HPCA68181.2026.11408569
Agent design, resource use, and Energy

05 **LLM INFERENCE ENERGY | NAACL 2025**
Poddar et al., Towards Sustainable NLP
aclanthology.org/2025.naacl-long.632
Output length, response time, and inference Energy

06 **MEASUREMENT STANDARD**
SCI for AI + ISO/IEC 21031:2024
greensoftware.foundation/standards/sci-ai
Functional unit and carbon-intensity method

Three actions to make every AI Agent step earn its place.

01 **Map every Model call, Tool call, Retry, and Handoff.**

02 **Remove or simplify work that does not change the accepted result.**

03 **Verify the change: keep the success check fixed, compare Trace workload, then measure Energy.**

Q&A

Questions or comments?

After the session: LinkedIn

GREEN AGENT DESIGN STARTS BY REMOVING UNNECESSARY WORK

We measured workload—not Energy or CO_2.

Can Workflow control reduce scheduled work while required outcomes stay fixed?

MEASURED DIRECTLY

- Scheduled Model-interface calls
- Exact serialized Model input size
- Planning / Reflection / Retry
- Retrieval documents and characters
- Tool schemas and actual Tool calls
- Memory load / write / retention

NOT MEASURED

- Live-LLM answer quality
- Provider Token telemetry or billing
- Production Latency
- Model inference Energy
- Carbon emissions
- General performance of other Agents

Purpose: establish auditable workload evidence before measuring Energy and CO_2.

The experiment creates auditable workload evidence before direct Energy measurement.

Repository guidance informed a benchmark built for this talk.

REUSED / ADAPTED

Repository guidance:
- Context scoping
- Model right-sizing
- Conditional Reflection
- Gated Retrieval
- Scoped Tools
- Memory + bounded Retry

Token-proxy pattern from the repo demo

BUILT FOR THIS TALK

- Six-case controlled Fixture
- Always-on and Lean Workflows
- Deterministic Model stand-in
- Recorder + JSONL Trace
- Scorer + validation checks
- CSV / JSON / Markdown outputs
- Reproducibility tests

SKILL.md: design-review guidance, not the runtime meter.

Repository: navveenb/lean-agentic-ai | commit 10940a535fcd…

SKILL.md can guide a review; benchmark.py performs the measured execution and logging.

Both Workflows received the same fixture and six different tasks.

Same for both conditions

6 tasks | 4 Knowledge Base docs | 7 Tools | history | rules | deterministic stand-in

01 BUSINESS HOURS
Code gate

PASS: Mon–Fri / 09:00–17:00

02 TICKET ROUTING
Classification + Tool

PASS: BILLING + route_ticket

03 CONTACT UPDATE
Extraction + state

PASS: Profile updated

04 RETURN POLICY
Selective Retrieval

PASS: 30 days

05 SESSION SUMMARY
Bounded Memory

PASS: 7 required facts

06 SLA ESCALATION
Plan + Tools + Reflect

PASS: Escalate for 3 reasons

Six different tasks—not six repetitions of one task. Each policy starts from its own copy of the same state.

The policies use separate copies of the same starting state and run the cases in the same order.

The Workflow policy was the only changed condition.

CONTROL	ALWAYS-ON	LEAN
Planning	Tier 3 for every case	Only complex SLA case
Model	Tier 3 for every call	Code / Tier 1 / Tier 2 / Tier 3
Retrieval	All 4 documents every case	One relevant document in 2 cases
Context	Full history + all evidence	Only task-required context
Tools	All 7 schemas every call	Only task-scoped schemas
Reflection	Every case	Only predeclared low-confidence SLA
Retry + Memory	Free-form retry; full transcript	Structured check; facts with TTL

Fixture = fixed test data | Tier = preassigned Model class | TTL = retention period

Tier, task type, and document key follow predeclared rules; the Agent did not learn them during the run.

Lean can remove, simplify, or retain work depending on the task.

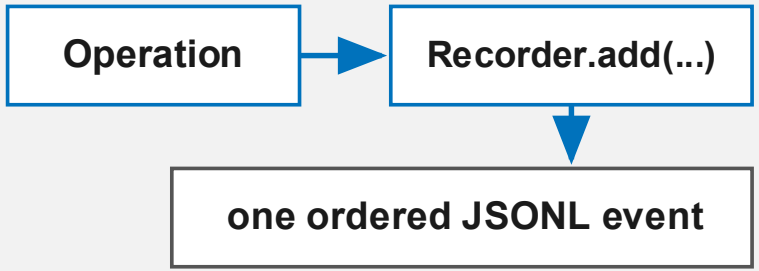
REMOVE	SIMPLIFY	RETAIN
BUSINESS HOURS	**CONTACT UPDATE**	**SLA ESCALATION**
Fixed rule → code only	Structured extraction + update	Plan + 3 Tools + Reflection
Model calls 3 → 0	Model calls 4 → 1	Model calls 3 → 3
PASS: same hours	Retry 1 → 0; profile PASS	PASS: required reasons

Lean means removing avoidable work—not removing work indiscriminately.

Necessary work remains; only work that does not help the accepted result is a removal candidate.

Recorder writes one Trace event after every operation.

AUTOMATIC RECORDING



Events: llm_call · retrieval · tool_call
memory · validation · case_complete

ONE llm_call EVENT (SHORTENED)

```
{ "case_id": "02-route",
 "stage": "answer",
 "model_tier": "tier1",
 "input_chars": 836,
 "input_token_proxy": 209,
 "exposed_tool_schemas": 1,
 "input_sha256": "…" }
```

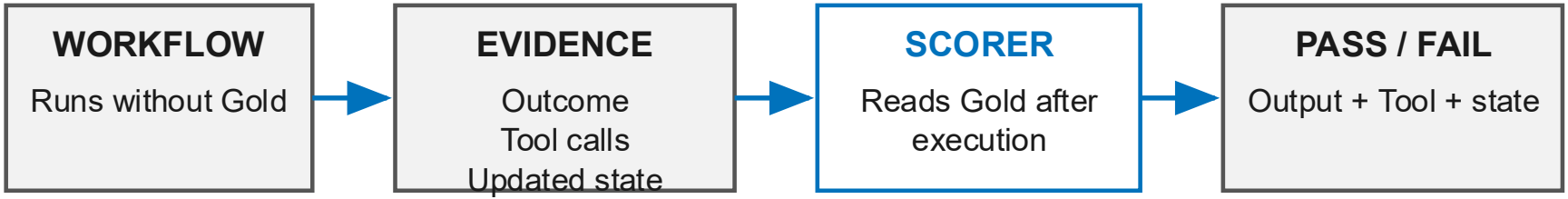
Token proxy = ceil(serialized characters / 4)

Exact stable JSON input and SHA-256 are retained; this is not provider Token telemetry.

Exact serialized input is retained; the Token proxy is a reproducible estimate, not Provider telemetry.

Scorer checks the business result only after execution.

Gold = the prewritten expected result; both Workflows run without it



ACCEPTANCE RESULTS

Structured outcome	Always-on 6/6 Lean 6/6	**Required actions**	Always-on 3/3 Lean 3/3	**State**	Always-on 1/1 Lean 1/1

Passing these checks does not establish equal natural-language quality.

Both conditions passed the same structured checks; natural-language quality was not evaluated.

Detailed metrics explain the 19 → 7 result.

MODEL CALLS BY CASE

	Always-on	Lean
Hours	3	0
Routing	3	1
Contact	4	1
Return policy	3	1
Summary	3	1
SLA	3	3
TOTAL	19	7

AGGREGATE WORKLOAD

	Always-on	Lean
Input token proxy	17,522	2,137
Planning	6	1
Reflection	6	1
Retry	1	0
Retrieved documents	24	2
Tool-schema exposures	133	8
Actual Tool calls	10	5
Memory loaded chars	4,160	262

The controls changed together; the observed reduction cannot be attributed to one control alone.

Repeated runs produced identical experiment outputs.

REPRODUCIBILITY CHECKS

- ✓ Same-order repeat: identical hashes
- ✓ Lean-first reversed order: identical
- ✓ Separate Python process: identical
- ✓ Different hash seeds: identical bytes
- ✓ Mechanical validation: PASS
- ✓ Unit tests: PASS

RUN + INSPECT

```
python3 benchmark.py
python3 -m unittest -v test_benchmark.py
```

```
PROTOCOL.md + fixture.json
benchmark.py
results/traces.jsonl
results/case_results.csv
results/validation.json
results/reproducibility.json
```

Deterministic totals—not six-run averages | Python 3.10.12 | standard library | external APIs: 0

The implementation uses Python 3.10.12, the standard library, and zero external API calls.

The result supports a Workflow claim—not an Energy claim.

SUPPORTED CONCLUSION

For this controlled six-task workload, conditional Workflow control reduced scheduled calls and serialized context while both conditions passed the same structured business checks.

NOT ESTABLISHED

- General reduction rate for other Agents
- Industry-average Always-on baseline
- Equal natural-language quality
- Intent-classification accuracy
- Retrieval-selection accuracy
- Individual effect of each Lean control
- TTL expiry or capacity eviction behavior
- Energy or CO_2 reduction

Workload reduction is evidence for the next measurement step—not an Energy claim by itself.

Workload reduction is the evidence needed before measuring Energy and CO_2 directly.