12

THE STACK BEHIND THE AI COWORKER

The Cheapest Token Is the One You Never Generate

Dr Peter McCann Strain, CTO and senior AI engineer, DPhil/PhD in AI from Oxford University

An agent workflow cut its token bill 71% with no new model, just by deciding who was allowed to decide.

An essay in the series **Architecting the AI Coworker**.
Approx. 23 minute read · Essay 12 of 22

Dr Peter McCann Strain
CTO, DPhil/PhD in AI from Oxford University

> The bill does not arrive as a finance line. It arrives as a trace, the step-by-step log of what an agent did. I was once handed the execution log of an agent workflow that everyone agreed was too expensive, and asked to find the waste. I expected one greedy step. Instead I found a relay. And every line of that relay was billed in the only unit these systems charge by: the token, the word fragment a model reads and writes, paid for by the one.

Before, the workflow was a sketch on a whiteboard: a support ticket goes in, an answer comes out. After eight months of patching, the trace told a different story. One support ticket came in. One model classified it. A second model summarised the classification. A third rewrote the summary. A fourth checked the rewrite. Then the whole thing was routed back to the first model, because by then nobody trusted the handoffs in between. The reasoning effort, the amount of internal deliberation each model is allowed before it answers, had crept up at every step. Routing had never been designed. A task that needed one coherent line of thought had quietly become five agents passing each other summaries.

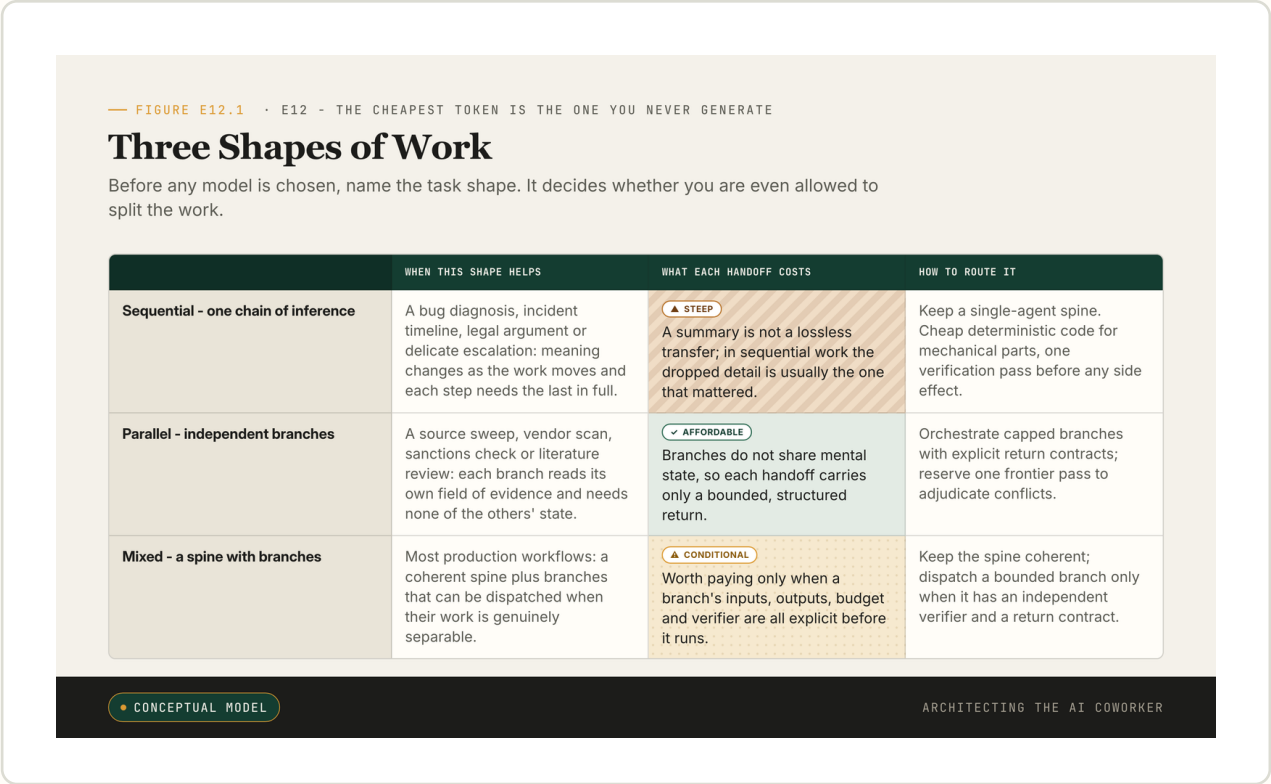
Here is what struck me. Every single decision in that chain was defensible on its own. The summariser was added because the classifier's output was messy. The checker was added because the rewrite sometimes drifted. The loop back was added the week a checker missed a refund error. Each addition solved a local problem on the day it was made. Together they built a system that spent money because nobody had decided how much authority each step deserved. The fix, when it came, was not a cleverer model. It was one coherent agent holding the ticket from first read to final reply, with cheap deterministic code for the mechanical parts and a single verification pass before anything was sent. The bill fell because the authority design changed, not because anyone shopped for a discount.

This essay is about that missing decision. The previous essay ended on a promise: that every runtime defence, every gate and route and extra check, spends budget, and that cost is the budgetary form of delegated authority. This essay makes good on it. It is about three dials almost no team consciously sets, and a discipline for setting them deliberately rather than letting them drift.

First, a distinction, because not all token waste is authority-design waste. A great deal of spend leaks out of plain systems inefficiency: poor retrieval chunking that pulls in three times the context a step needs, missing caching that pays again for the same prompt prefix, duplicate context stuffed into every call, sloppy prompt templates, retries that mask a flaky tool, vendor pricing that quietly changes under you. That waste is real and worth chasing, but it is an engineering hygiene problem with engineering hygiene fixes. This essay is about the other kind: the waste that comes from ungoverned delegation, from a workflow that never decided how many minds to trust, how hard each may think, or how many separate jobs the work should become. The authority frame is powerful, and for that reason it must not be asked to

absorb every cost-engineering question. Fix the plumbing first. Then ask the harder question this essay is for.

Routing is not only a cost control. It decides which model is allowed to make a judgement. Effort is not only a latency control. It decides how long the system may think before it acts. Agent count is not only a diagram on a whiteboard. It decides how many context boundaries, summaries, handoffs and small decisions the workflow will create. The lesson is not that cheap models are virtuous, or that multi-agent systems are wasteful. The lesson is narrower and more useful. Cost is the shadow cast by an authority design. The cheapest token is the one the system never needed permission to generate.



Three shapes of work decide whether you may split a task, before any model is chosen.

Three dials drift, every time, towards maximum spend

Most agent spending is not approved. It accumulates. The relay above is the long version of the story; the short version is that the control surface has no name, so nobody guards it.

It has three dials, and each one wears the wrong label. Start with routing. The router, the component that decides which model handles a given request, presents itself as a cost knob, a choice about price and capability. It is not a cost knob. It is an authority knob: it decides which mind is trusted with the judgement. Effort wears a second false label. The amount of internal deliberation a model is allowed before it answers passes for a latency control, a dial about how fast the answer comes back, when what it really sets is how long the system gets to think before it commits. And agent count, which everyone reads as an architecture choice about decomposition and parallelism, is quietly deciding how many places in the workflow partial un-

derstanding will be allowed to look complete, because every extra agent is another context boundary and another delegated authority.

Treat those three as separate, unrelated decisions and the system drifts, every time, towards the same place: frontier model, full effort, more agents. Treat them as one decision and the question changes. What kind of work is this. What can verify it. What happens if it is wrong.

It helps to write the bill as a small equation, because each term maps to a dial. The cost per run is the spine tokens, the work of the one coherent line of reasoning, plus the branch tokens, everything spent in agents split off the spine, plus retries, plus verification, the cost of whatever checks the output, plus the human queue cost, the salaried time a person spends when the workflow hands a decision back. The five-model relay was expensive because the second and third terms had swollen with no one watching, and the last term, the loop back, had been bolted on to compensate. An authority design is, in the end, a set of decisions about how large each of those five terms is allowed to grow.

Table 1. The relay, scored against one ticket (composed bill with one cited real-trace anchor). Most rows are composed, not measured, and should be replaced with your own traces before any procurement decision; they are here to make each term in the equation concrete. The branch / handoff row is anchored to a published figure. Anthropic's engineering write-up on its multi-agent research system reports that the multi-agent design uses roughly fifteen times more tokens than a single chat with the same model on the same task; that is the operational order of magnitude for the branch-and-handoff overhead a workflow drags in once it is split [5]. The percentages in the right-hand column are stipulated reductions on the composed bill except where the row is flagged "real-trace anchor". And the spine row deliberately rises: the single coherent agent costs more per run because it now holds, in one context, the reasoning the four-way chain used to fragment. The work moved onto the spine, it did not vanish, which is exactly why the branch, retry and verification terms collapse and the total still falls.

TERM IN THE EQUATION	OLD DESIGN (V0)	REDESIGNED (V1)	TOKEN CHANGE	SOURCE
Spine tokens (one coherent agent does more)	~4,000 (4 models in chain)	~9,000 (1 model)	+125% (the spine absorbs work the relay used to scatter)	composed
Branch / handoff context (agent count)	~15x single-agent baseline	1x baseline (single coherent agent)	-93% on this term	real-trace anchor: Anthropic, "How we built our multi-agent research system" [5]

TERM IN THE EQUATION	OLD DESIGN (V0)	REDESIGNED (V1)	TOKEN CHANGE	SOURCE
Retries and re-classification	~14,000 amortised	0 (loop removed)	n/a	composed
Verification (reasoning effort)	bundled into branches (medium each)	~2,000 (low + one explicit pass)	-52%	composed
Human queue cost	~3 minutes per 1 in 5 tickets	~3 minutes per 1 in 5 tickets (unchanged)	n/a	composed
Total tokens / ticket (avg)	**~38,000**	**~11,000**	**-71%**	composed

Most rows in Table 1 are stipulated, not benchmarked. Only the branch / handoff row is anchored to a real published figure (the ~15x above, from Anthropic's own write-up [5]); read the rest as a worked illustration of the equation, not a target. The point is the structure: each term maps to one dial, so each dial has a place to show up on the bill.

What shape is the work, and may you split it?

Before any model is chosen, the prior question is what kind of work this is. There are three shapes, and the distinction is not academic. It decides whether you are even allowed to split the task.

Sequential work is one chain of inference. A bug diagnosis, an incident timeline, a legal argument, a migration plan, a delicate customer escalation: the meaning of the work changes as it moves. The second step depends on the first step's full context, not a summary of it. The third step depends on the second step's unresolved doubts. A summary is not a lossless transfer, and in sequential work the thing a summary drops is usually the thing that mattered.

Parallel work is a set of independent branches. A source sweep, a vendor scan, a sanctions check, a literature review, a competitive analysis: each can send a different worker into a different field of evidence, because one branch does not need the whole mental state of the others to do its job.

Mixed work has a spine and branches, and most production workflows live here. The discipline is to keep the spine coherent and dispatch a bounded branch only when its inputs, outputs, budget and verifier are all explicit.

The empirical support for this is still fast-moving, much of it in preprints, work published before formal peer review, and I want to use it with appropriate care. A Google and MIT scaling study

(preprint) describes a large body of configurations across multiple benchmarks and reports a strong positive result on decomposable work, while the same line of evidence warns that sequential tasks can lose ground once multi-agent coordination adds its overhead ([3], preprint). Tran and Kiela (preprint) sharpen the mechanism: under equal thinking-token budgets on multi-hop reasoning tasks, the kind that require chaining several facts together, the handoffs between agents act as information bottlenecks, so a single agent with coherent context can match or beat multi-agent designs in that tested setting ([4], preprint).

That gives the operating rule worth carrying through the whole essay: every handoff is a compression event. When work passes from one agent to another, what crosses the gap is a summary, and a summary has dropped something.

Sometimes the compression is worth paying for. Anthropic's account of its own multi-agent research system is a serious example. It uses breadth-first research, meaning the workflow fans out to explore many leads at once rather than following a single chain deep, with an orchestrator, a coordinating agent that dispatches the workers and assembles what they return, and a task family where parallel exploration buys useful coverage. Anthropic is candid in the same write-up that this design consumes substantially more tokens than a single-agent chat, and treats that as a cost the task has to justify [5]. I should be plain that this is a vendor engineering post, not independent measurement, so it is best read as an honest worked example rather than a benchmark. The lesson stands either way. Breadth is not free, and it is not always intelligence.

A router is an automated decision a regulator can audit

There is something the cost frame can hide, and it is worth marking before the rubric. When a router decides which model is allowed to make a given judgement, that routing decision is itself an automated decision, and several jurisdictions have started to treat it as one.

The plain operating rule, drawn from where the regulation already bites, runs like this. In California, when automated decision-making technology drives a significant decision about someone's job, housing, credit, insurance, healthcare or schooling, the organisation owes that person notice before the fact, a way to see what happened, and a way to opt out [6]. Quebec comes at the same problem from the individual's side: where a decision is made by automated processing alone, the person is entitled to be told the main factors behind it and to put their case to a human reviewer [7]. Europe attacks it through the record, requiring high-risk systems to log automatically, enough to trace how the system behaved across its life [8].

That is three jurisdictions converging from different directions, and the prevailing US risk framework simply adds the housekeeping rule beneath them all: it expects you to have written the third-party-model question down before anyone asks [9]. One operative regime, three corroborating doctrines, and a single conclusion none of them dodges: an automated routing decision is a documentable one.

The operational point is plain. A router that sends a credit decision, a hiring screen, or a clinical triage to one model on Monday and a different model on Tuesday is not making a private engineering choice. It is making a regulated one. The five fields the router should log, the shape, the tier, the effort, the verifier and the handoff condition, set out in full when we reach the rubric, are exactly the fields that turn a routing decision from a private cost optimisation into an auditable record under those regimes. Treat them that way from the start, and the audit trail is a by-product of the cost-engineering work rather than a retrofit demanded after the first complaint.

What does enforcement against an undocumented router actually look like? The rules are mostly too new to say. California's regime only took effect at the start of 2026, so the first formal actions for logging non-compliance are still working through the pipeline, and the nearest live signal is Canadian. In a September 2025 joint investigation of TikTok, the federal privacy commissioner and provincial counterparts found the company's collection and use of children's personal information inappropriate, faulting its ineffective age checks and its inadequate consent and transparency, and ordered remediation [11]. That case did not turn on documented governance over the automated systems themselves. But the gap it exposed, the absence of a record explaining why a system decided as it did, is exactly the one the five logged fields are meant to close. A router that cannot produce them is one complaint away from a comparable documentation finding.

Worked mini-case: one relay, redesigned

The same relay, told term by term. The old design ran five model calls in series (classifier, retriever, drafter, policy-checker, sentiment-trimmer) on a single frontier tier, with handoffs repeating context and a re-classification loop firing on roughly one ticket in three. The redesigned workflow runs a single-spine drafter on a mid-tier model with one explicit verifier pass and a tighter handoff condition; the human-handoff rate is held at roughly one in five, because the policy bar did not move. Three things are pinned across the comparison, and naming them is what keeps it honest: the same escaped-error bar, the same human-handoff rate, the same policy threshold. So the 71% is a pure shape dividend, not quality quietly sold off to buy savings; a hostile reader cannot dismiss it as moved goalposts, because the goalposts are listed. The relay was not a model problem. It was a shape problem priced as a model problem.

Two pieces of public research mark out the stakes, and both are preprints, work circulated before formal peer review, so they are read here as direction rather than settled measurement. CLEAR (preprint) refuses to grade agent systems on accuracy alone; its evaluation frame is Cost, Latency, Efficacy, Assurance and Reliability, and its central finding is that pushing for a sliver of extra accuracy can drive cost up sharply while doing nothing measurable for operational risk ([1], preprint). RouteLLM (preprint) shows the opposite posture paying off: route the easier requests down to a cheaper model when the quality target still permits it, rather than making every request pay the frontier tax, where a frontier model means one of the largest and

most capable models a vendor offers, priced accordingly ([2], preprint). Under all this sits a published floor. Park et al.'s peer-reviewed Generative Agents work documents what a multi-agent simulation actually costs in time and tokens once roles, memory and reflection are separated, giving the empirical spine a reference point the preprints do not have to carry alone [10]. And on the routing side specifically, a harder number: Chen, Zaharia and Zou's FrugalGPT, peer-reviewed in late 2024, reports an adaptive LLM-cascade framework matching GPT-4 quality at up to 98 percent cost reduction across natural-language tasks, so the route-down-when-possible posture rests on a peer-reviewed result, not only a preprint [12]. The strongest agent architecture is not the one that thinks hardest everywhere. It is the one that can say where thinking is not the scarce resource.

Treat every default as a position you must argue out of

This is the artefact. Use it before deployment, and again whenever the traces show high cost, slow latency, repeated retries or unclear ownership.

DECISION	DEFAULT	ESCALATE WHEN	VERIFICATION TRIGGER	HUMAN HANDOFF TRIGGER
Task shape	Single-agent spine.	Branches have independent inputs and can return structured evidence.	Branch output must cite evidence, confidence and what it did not inspect.	Branches conflict on a load-bearing fact or cannot explain their basis.
Routing tier	Cheap model or deterministic code.	Ambiguity is real, confidence is low, or the output will shape a later judgement.	Medium/frontier output must be checked by rules, source comparison, or a second model with a different prompt.	The decision is irreversible, legal/regulatory, financial, safety-relevant, or externally visible.
Effort budget	Zero or low reasoning.	The answer is load-bearing, hard to verify mechanically, or likely to trigger an external action.	Full-effort work must leave a reasoned decision record and an explicit uncertainty field.	Additional effort no longer changes the decision, or the model cannot resolve a material conflict.

DECISION	DEFAULT	ESCALATE WHEN	VERIFICATION TRIGGER	HUMAN HANDOFF TRIGGER
Agent count	No new agent.	The branch is parallel, capped, independently checkable, and has a return contract.	Each agent must return only the fields needed by the spine.	More agents are creating summaries of summaries, conflicting assumptions, or unbounded tool use.

The rubric is not anti-frontier. It is anti-default. Read each row as a starting position you must argue your way out of, not into.

The cheap tier, the smallest production-grade models, handles classification, extraction, deduplication, formatting, routing and other work where correctness can be checked quickly and cheaply. The medium tier handles comparison, synthesis, reconciling competing drafts and mapping uncertainty. The frontier tier, the strongest and most expensive models, is reserved for high-consequence ambiguity, the cases where the judgement itself is the hard part and getting it wrong is costly. Effort follows the same logic: zero or low reasoning by default, full reasoning only when the answer is load-bearing and hard to verify by mechanical means.

Agent count is governed by the same principle, and it is the dial teams get wrong most often. Do not start with "three agents". Start with the spine, one coherent line of work, and then ask of each candidate branch whether it deserves its own budget, its own context, its own permissions, its own verifier and a return contract: an explicit statement, written before the branch runs, of exactly which fields it must hand back to the spine and in what form. If a branch cannot be verified independently of the spine, and cannot be given a clean return contract, it is probably not a branch at all. It is a paragraph inside the same reasoning chain that someone mistook for a separate job.

A decision-grade router needs thresholds, not instincts. Route down when the task is reversible, low-consequence, mechanically checkable, or already pinned by a fixed schema, and route across when several independent fields of evidence have to be read at once. Reserve routing up for the answers that are ambiguous, consequential, hard to verify with rules, or likely to fire an external action. And stop entirely, handing to a person, when the model cannot resolve a material contradiction, when the tool action is irreversible, when a live legal or regulatory judgement is needed, or when more effort is only producing different words rather than better evidence. Those conditions overlap, so the router needs a tie-break for when two fire at once. The rule is precedence by consequence: when a route-down condition (reversible, low-stakes) and a route-up condition (consequential, hard to verify) both apply to the same task, consequence wins and you route up. Cost is recoverable; an irreversible wrong judgement is not. Two readers handed the same boundary case will otherwise route it two different ways, and the instrument stops being reproducible.

And the router should log five fields for every non-trivial escalation: the shape, which explains why the work stayed in one spine or split into branches; the tier, which explains why the chosen model was enough; the effort, which explains why the system spent more or less deliberation; the verifier, which explains what checked the output; and the handoff condition, which explains when automation must stop. Those are the five fields the audit regimes were reaching for. They make cost review possible, and they make governance possible at the same time. If a workflow cannot explain why a step used a frontier model, why it spawned a second agent, or why it acted without human review, then the problem was never the bill. The bill is just the first audit trail that happened to arrive.

Run the rubric on the workflow you started with

Go back to the five-model relay from the start of this essay. Run that workflow down the rubric and the diagnosis is immediate. Its task shape is a sequential spine: triaging a ticket is one chain of inference, where the right reply depends on the full sense of the complaint, not a summary of a summary of it. So the default, single-agent spine, was the correct answer, and the team argued its way into four extra agents without ever arguing its way out of the default. Each handoff was a compression event, and the loop back to the classifier was the system itself noticing that compression had cost it the thread. In the cost equation, the relay had let the branch term and the human queue term run while no one priced them; collapsing it back to a spine shrank both at once.

That is one architecture the control surface produces. It is worth setting a second beside it, to show the surface does not always say "do not split". Consider a supplier-risk scan: checking a new vendor against sanctions lists, financial health, security posture, contractual exposure and operational dependency. That work is parallel. Each of those five checks reads a different field of evidence and does not need the others' full mental state to do its job. The wrong design is one frontier agent reading everything in a single long context, or an uncapped swarm exploring freely and returning prose. The right design is an orchestrator that creates five capped branches, gives cheap or medium workers a clear return contract for each, and reserves one frontier synthesis pass to adjudicate conflicts, with a human handoff if sanctions or legal exposure stays unresolved.

Set the two side by side and the symmetry is the point. The relay gets safer by refusing to split the thinking. The supplier scan gets safer once you split the search but keep the judgement in one place. What is scarce in the relay is continuity; for the scan it is coverage. Same three dials, set to opposite values, because the shapes pull in opposite directions. None of this is a benchmark claim. It is the rubric applied honestly to two shapes of work, resting on directional evidence: that accuracy-only optimisation can drive cost up [1], that routing can cut cost while holding target quality in evaluated settings [2], and that task shape changes whether decomposition helps at all [3][4].

It is worth being precise here about how far that evidence actually carries, before the frame is asked to bear any weight. CLEAR supplies the cost-aware evaluation frame ([1], preprint),

RouteLLM the routing result ([2], preprint), and the Google/MIT and Tran/Kiela papers the task-shape and handoff-compression mechanism ([3], preprint; [4], preprint); all four are preprints, and a preprint can shift between circulation and formal publication. Beneath the preprint line sits a peer-reviewed floor (Park et al. on what multi-agent simulation costs in time and tokens; FrugalGPT on cost-quality routing cascades), and beside it a candid vendor case (Anthropic) that is honest about its own token cost but audited by no one outside the company. The preprint line is direction; the peer-reviewed line is the measured floor. What that body of work does not settle is three things:

1. Where your own workflow sits on the sequential-to-parallel spectrum. That answer lives in your traces and in no paper.
2. Whether to tune your router by confidence thresholds, cost ceilings, latency ceilings, action class or some blend. That turns on a harm model only your team can write down.
3. How long any current preprint number will survive contact with the next model release.

Those gaps are real. They are not reasons to wait. The mature architecture is therefore not a clever router. It is a router with an audit trail.

So far this is two worked architectures and the evidence behind them. Now the objection the frame would be dishonest to skirt, and it happens to be correct. Heterogeneous multi-agent systems, teams of differently specialised agents, can outperform a single agent when the roles are specialised, the branches are independent, and the outputs are externally verifiable. That is precisely why this argument must not curdle into a cheap-model manifesto. Anthropic's research system is a worked example of multi-agent breadth helping, in a workload built for search, coverage and synthesis [5]; the Google and MIT scaling study supports a task-shape view of the question rather than an agent-count ideology; and Tran and Kiela's equal-budget critique does not prove that one agent always wins, only that handoffs have to earn their keep when the work is sequential and context-sensitive [3][4]. So the answer is not "never split". The answer is "split only what can be returned". A branch should hand back evidence, not vibes; fields, not a mood; conflicts laid bare, not a polished paragraph that hides its own uncertainty.

A sharper version of the same objection turns the cost frame on itself. Every defence in this essay, the explicit verifier pass, the return contract, the five logged fields, is itself token spend and delegated authority by my own definition, so why is any of it exempt from the scrutiny I aim at the relay? Who verifies the verifier? The frame answers on its own terms: a verifier earns its budget only when its check is cheaper and more reliable than the failure it catches. When verification would cost more than the error it prevents, the honest move is not to stack another check on top; it is to route up or hand to a person. The recursion stops where a check stops paying for itself.

That is the difference between delegation and diffusion, and it is the heart of this essay. Delegation names the authority, the scope, the budget, the verifier and the return contract. The branch knows what it is allowed to do and what it must hand back. Diffusion simply creates more places where partial understanding can dress itself up as a complete answer. A multi-

agent system can be either. The diagram looks identical from a distance. The difference is whether anyone wrote the contracts.

So here is the move for Monday morning. Take your most expensive recurring agent workflow and write one row per sub-task: shape, routing tier, effort budget, verifier, human handoff condition. To make the artefact reproducible rather than a matter of instinct, populate one row fully before you judge the rest. For the relay's classifier sub-task a finished row reads: shape = sequential-spine; tier = mid (frontier not earned, the output is checkable); effort = low; verifier = one explicit pass before send; handoff = any unresolved refund or legal flag. That is the threshold a second reader should be able to reproduce from the same trace. Then look down the rows and ask the uncomfortable question. Which row still says "frontier, full effort, separate agent" because the task earns it, and which row says it only because nobody ever changed the default? The row that earns nothing different from the default is the one to delete.

This is where the series threads meet. Tools give models hands, and supply chains govern which of those hands are ever admitted. Prompt injection, the subject of the last essay, is about whose sentences get to steer them; task shape, the subject of this one, settles how many models, tokens and delegated judgements the work actually deserves. And after spend comes memory. Once the agent has done the work, what should it remember, who can erase it, and which future action may that memory quietly influence? The next cost is not the token you generate. It is the state you keep.

Carry This Forward

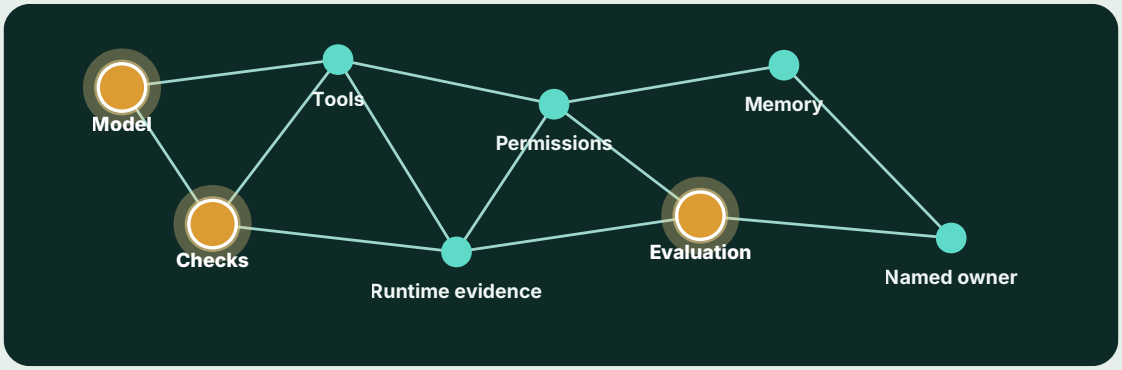
> A bill that runs high is not always a story about extravagance. It is often a story about a workflow that never decided who was allowed to decide: which model may judge, how much effort a step may spend, and how many agents get a delegated piece of the work.
>
> **One move:** Take your most expensive recurring workflow and write one row per sub-task: shape, routing tier, effort budget, verifier, human handoff condition. Every handoff compresses context, so name the task shape before you add a model call or a sub-agent. The row that earns nothing different from the default is the one to delete.
>
> **Next:** The next form of authority is memory, the state that survives the run. Once the work is done, the question is what the system keeps, who can erase it, and what that memory may influence later.

THE STACK SO FAR

E12 · Essay 12 of 22 complete · Arc III: Runtime control

The Stack So Far. *Every essay adds one instrument to the operating model. The constellation shows which eight you are building, which are lit by essays you have read, and which is added right here.*

I See the object | II Evidence and authority | III Runtime control — ESSAY 2 OF 4 | IV Proof and accountability | V Operating model



built in earlier essays · added in this essay · coming in later essays

01 02 03 04 05 | 06 07 08 09 10 | 11 12 13 14 | 15 16 17 18 19 | 20 21 22

You have just added.

The task-shape routing rubric

You can now route work by task shape before model choice.

Next. E13 asks what an agent remembers and what it cannot forget.

← PREVIOUS
E11 · The Sentence That Owns the Agent

Essay 12 of 22 complete

NEXT →
E13 · What an Agent Cannot Forget

References

Reference links for sources cited in this essay.

1. **Beyond Accuracy: A Multi-Dimensional Framework for Evaluating Enterprise Agentic AI Systems (CLEAR)**
   Sushant Mehta
   https://arxiv.org/abs/2511.14136

2. **RouteLLM: Learning to Route LLMs with Preference Data**
   Ong et al.
   https://arxiv.org/abs/2406.18665

3. **Towards a Science of Scaling Agent Systems**
   Kim et al. (Google Research, Google DeepMind, MIT)
   https://arxiv.org/abs/2512.08296

4. **Single-Agent LLMs Outperform Multi-Agent Systems on Multi-Hop Reasoning Under Equal Thinking Token Budgets**
   Dat Tran and Douwe Kiela
   https://arxiv.org/abs/2604.02460

5. **How we built our multi-agent research system**
   Anthropic
   https://www.anthropic.com/engineering/multi-agent-research-system

6. **CCPA updates: ADMT, risk assessments and cybersecurity audit regulations**
   California Privacy Protection Agency
   https://cppa.ca.gov/regulations/ccpa_updates.html

7. **Act respecting the protection of personal information in the private sector (Quebec Law 25), section 12.1**
   National Assembly of Quebec
   https://www.legisquebec.gouv.qc.ca/en/document/cs/P-39.1

8. **Regulation (EU) 2024/1689 (EU AI Act), Article 12: record-keeping / logging for high-risk AI systems**
   European Parliament and Council
   https://eur-lex.europa.eu/eli/reg/2024/1689/oj

9. **AI Risk Management Framework (AI RMF 1.0), NIST AI 100-1**
   National Institute of Standards and Technology
   https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf

10 **Generative Agents: Interactive Simulacra of Human Behavior**
Park et al.
https://dl.acm.org/doi/10.1145/3586183.3606763

11 **Joint investigation of TikTok Pte. Ltd. and TikTok Technology Canada Inc.: PIPEDA Report of Findings #2025-003**
Office of the Privacy Commissioner of Canada (with provincial counterparts)
https://www.priv.gc.ca/en/opc-actions-and-decisions/investigations/investigations-into-businesses/2025/pipeda-2025-003/

12 **FrugalGPT: How to Use Large Language Models While Reducing Cost and Improving Performance**
Chen, Zaharia and Zou
https://openreview.net/forum?id=cSimKw5p6R

About the Author

ARCHITECTING THE AI COWORKER

Dr Peter McCann Strain

Dr Peter McCann Strain is a CTO, founder and senior AI engineer with a DPhil/PhD in AI from Oxford University. He builds production AI systems and writes about making agentic AI useful, inspectable, governable and safe enough for real work.

READ THE FULL SERIES

Substack (canonical)	petermccannstrain.substack.com
Medium	@peter.mccann.strain
LinkedIn	peter-strain-dphil-15a607128
Web	petermccannstrain.com
Cadence	New essays twice weekly, 2 June – 21 July 2026