Multi-agent work can burn ~15x a single chat. Cost is an authority design.

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

Swipe ››

— THE BUYER RISK

A task that needed one coherent spine becomes five agents handing each other summaries. Routing, effort and agent count get treated as separate local decisions, so the system drifts towards frontier, full effort, more agents, and spends because it never decided how much authority each step deserves.

— THE REFRAME

Cost is authority design.

THE OLD QUESTION

Which model is cheapest?

↓ ↓ ↓

THE QUESTION THAT HOLDS UP

What is the task shape (sequential, parallel or mixed), and what authority (how much budget, effort and tool reach) does each step deserve?

— WHAT TO ASK FOR

~15x token spend

Anthropic reports its own multi-agent research system uses roughly 15x more tokens than a single chat on the same task. The right answer is not 'more agents'. It is matching task shape to authority. A real support-ticket relay fell from ~38,000 to ~11,000 tokens per ticket once redesigned as one coherent agent with deterministic checks (single case; most cost rows illustrative).

SOURCE

Anthropic, 'How we built our multi-agent research system' (13 Jun 2025), for the ~15x token figure; author's anonymised reference analysis for the relay example; external posture from Tran and Kiela's equal-budget critique, CLEAR's cost-aware evaluation frame, RouteLLM, and the Google/MIT scaling work.

— CHECKLIST LOGIC

Name the task shape before choosing a model.

01 Keep **sequential** work on one spine, because every handoff compresses context.

02 Split **parallel** work into branches that can return structured evidence.

03 Dispatch **mixed** branches only with explicit inputs, budget and verifier.

THE ARTIFACT

Same task, 38k tokens to 11k: task shape priced the bill.

Before

5 calls / repeated summaries / unclear owner

After

2 calls / one spine / one verifier

Saved

fewer tokens and fewer authority handoffs

Illustrative, n=1: a real five-model relay (classifier, retriever, drafter, policy-checker, sentiment-trimmer) fell from ~38,000 to ~11,000 tokens per ticket after the workflow shape changed. Single case with most rows composed, not a benchmarked distribution.

— ASK THIS ON MONDAY

Take your most expensive recurring agent workflow this week. Write one row per sub-task: shape, routing tier, effort budget, verifier, handoff condition. Flag every row that says 'frontier, full effort, separate agent' only because nobody changed the default.

VENDOR TRAP

Adding a cheaper model to cut the bill. Cost lives in task shape, not unit price. Redesign the spine before swapping the model, or you will pay frontier rates on a relay that did not need five agents.

— USE THE CHECKLIST

The Cheapest Token Is the One You Never Generate

Read the full essay — the argument, the sources, the figures and a reader-ready working artifact.
Substack petermccannstrain.substack.com · Medium @peter.mccann.strain · LinkedIn peter-strain-dphil-15a607128
New essays twice weekly, 2 June – 21 July 2026.

Next: E13 — What an Agent Cannot Forget

— THE STACK SO FAR

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

YOU JUST ADDED

The task-shape routing rubric

STACK LAYER LIT UP

Model / Checks / Evaluation

YOU CAN NOW ASK

route work by task shape before model choice.

NEXT

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

— THE ARTIFACT, CONTINUED

Same task, 38k tokens to 11k: task shape priced the bill.

THE REMAINING NODES

Row template (one per sub-task): shape | routing tier | effort budget | verifier | handoff condition.

Shape: sequential, parallel or mixed.

Routing tier: frontier, mid, small (the cheapest tier that meets the bar).

Effort budget: token or call ceiling before the step must yield.

Verifier: who or what confirms the step before handoff.

Handoff condition: the named trigger that returns control.

Dr Peter McCann Strain

CTO, DPhil/PhD in AI from Oxford University

I build production AI systems and write about making agentic AI useful, inspectable, governable and safe enough for real work.

Follow on Substack for the full 22-essay series
petermccannstrain.substack.com