CLAUDE / MANAGED AGENTS

Plan big.
Execute small.

A practical Fable 5 + Sonnet 5 orchestration workflow for expensive, multi-step Claude tasks.

96% **of all-Fable performance**
at about 46% of the cost on Anthropic's BrowseComp test

- **Best AI Tools I Use Daily**
  https://tools.johnsavageai.com/
- **All My Socials**
  https://linktr.ee/johnsavage14

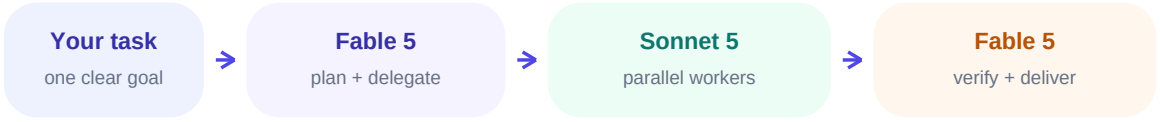
THE IDEA

Let the strongest model think, route, and check the work. Let the cheaper model handle the high-volume execution leg.

Fable 5 stays focused on planning and judgment; Sonnet 5 workers take on repeatable, well-scoped tasks in parallel, then Fable 5 synthesizes the result.

THE LOOP

Plan -> delegate -> execute -> synthesize



USE THE RIGHT EXPECTATION

The 46% figure is a benchmarked cost result, not a promise of 46% fewer tokens. Savings vary with task shape, delegation overhead, and tool use.

01 / SETUP

Set it up in Claude

Open Claude, enable the tools, then give it the cookbook and the instruction.

Open Claude

https://claude.ai

1 **Open Claude**
Start a new chat, Project, or workspace where you want the workflow to live.

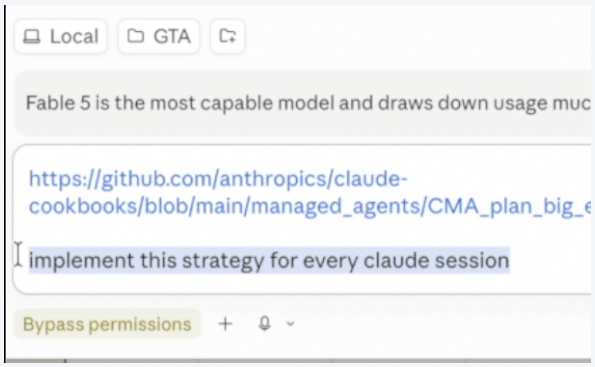
2 **Enable the tools**
Turn on Computer Use and Browser Use in the tools or permissions menu.

3 **Paste the cookbook**
Paste the GitHub link below so Claude can inspect the managed-agents method.

4 **Add the instruction**
Paste the copy-ready prompt on the next page after the link.

Reference screenshots

The visual pattern is simple: link first, instruction second.



Paste the GitHub link, then add the instruction.

The reference wording is: *implement this strategy for every Claude session.* The copy-ready version below makes the model split explicit and adds the safer cost-performance wording.

https://github.com/anthropics/claude-cookbooks/blob/main/managed_agents/CMA_plan_big_execute_small.ipynb

Use the full URL so Claude can open the exact notebook.

WANT THIS TO PERSIST?

Save the prompt as a Project instruction or persistent workspace instruction if your Claude setup offers it. A normal chat only affects the current session.

02 / PROMPT

Copy this prompt

Paste it after the GitHub link in Claude.

COPY-READY PROMPT

```
https://github.com/anthropics/claude-cookbooks/blob/main/managed_agents/CMA_plan_big_execute_small.ipynb

Implement this strategy for every Claude session.

Use Claude Fable 5 as the orchestrator for planning, delegation, and final synthesis.
Delegate token-heavy, well-scoped execution work to Claude Sonnet 5 workers, using
parallel delegation whenever subtasks are independent. Have Fable 5 verify and
synthesize the final result.

When the task needs it, use Computer Use and Browser Use. On suitable workloads,
aim for roughly 96% of all-Fable performance at about 46% of the cost. Treat this
as a benchmark, not a guarantee.
```

GOOD FITS

Research, codebase sweeps, document review, log analysis, repetitive web tasks, and any workflow where a lot of reading or mechanical execution can be split into smaller jobs.

Useful links

- **Claude**
  https://claude.ai
- **Managed Agents docs**
  https://platform.claude.com/docs/en/managed-agents/multi-agent
- **GitHub cookbook**
  https://github.com/anthropics/claude-cookbooks/blob/main/managed_agents/CMA_plan_big_execute_small.ipynb

Built from Anthropic's Multi-agent sessions documentation and the Plan big, execute small cookbook. The reported 96% / 46% result is an Anthropic benchmark reference; rerun your own tasks to measure the real outcome.