Run 100 agents at once in Claude Code

Here's how to do what I showed in my video. In Claude Code, you put the word `ultracode` *in a prompt, and Claude writes a small program that runs many subagents for you. You can ask for 100. You won't write any code.*

What you need

- A paid Claude plan. This feature, called dynamic workflows, is on all paid plans. On Pro, turn it on first: type `/config` and switch on the **Dynamic workflows** row.
- **Claude Code**, not the regular chat, kept up to date.
- A folder of real work to point it at, such as a project with many files.

Step 1: ask for a workflow

Type `ultracode:` and then your task, in Claude Code:

```
ultracode: audit every API endpoint under
src/routes/ for missing auth checks
```

Claude Code highlights the keyword, and Claude writes a workflow for the task instead of working through it one step at a time. The keyword only changes how Claude organizes the work. The agents get the same permission checks as anything else in your session.

Asking in your own words (“use a workflow”, “run a workflow”) also works.

If you didn’t mean to start one, press `Option+W` on a Mac or `Alt+W` on Windows to dismiss the highlight for that prompt.

Step 2: say how big the team is

A subagent is a helper that takes one piece of the job. Claude aims for a team size set in `/config`

under Dynamic workflow size:

Setting	Claude aims for
small	fewer than 5 agents
medium	fewer than 10 agents
large	fewer than 50 agents
unrestricted	whatever fits the task

The default is medium, or small if you're on Pro. This is advice, not a limit. A prompt that asks for a different size overrides it, so you can write:

```
ultracode: spin up 100 subagents and review every file in this repo for bugs, then give me one report
```

Step 3: approve the plan

Before it runs, Claude Code shows the plan with its phases. In the terminal, choose **Yes, run it**, **View raw script** to read what Claude wrote, or **No**. In the desktop app you get a card with **Once**, **Always** and **Deny**.

In auto mode, you're asked only on the first launch, and not at all if you've turned on `/effort ultracode` for the session. In bypass permissions mode, it starts without asking.

Step 4: let it run, and watch

It runs in the background, so your session stays usable. Type `/workflows` to see each phase, how many agents are in it, and each agent's token use. Press `p` there to pause the run, or `x` to stop it.

When it finishes, the result comes back into your session as one report.

How big a run can get

- **Up to 1,000 agents in one run.** 100 is well inside that.
- **16 run at the same time by default,** fewer if your computer has fewer CPUs. The rest wait their turn. You can change that with the `CLAUDE_CODE_WORKFLOW_MAX_CONCURRENT_AGENTS` setting, from 1 to 256, on Claude Code v2.1.269 or later.

- **You can’t type into a run halfway.** It pauses by itself only for an agent’s permission prompt or when you hit a usage limit. If you want a sign-off between stages, run each stage as its own workflow.

Keep it inside your plan

A big run uses a lot of your plan. Runs count toward your usage limits.

- **Start on a small slice,** one folder instead of the whole repo, and watch the token numbers in `/workflows`. You can stop the run there, usually without losing the work already done.
- **Look out for the** `Large workflow` **warning.** It shows when a run goes past 25 agents or about 1.5 million tokens. It only warns: it doesn’t stop anything.
- **Check** `/model` **first** if you normally switch to a smaller model for routine work. Agents use your session’s model unless told otherwise. You can ask Claude to use a smaller model for the easy stages.

- **If you hit your usage limit,** the run pauses and carries on by itself after the reset, as long as the app stays open. This needs a recent version (Claude Code v2.1.271, or v2.1.319 in the desktop app's Code tab), a subscription without usage-based billing, and a limit that resets within 24 hours. It waits at most twice. After that, agents fail, and you can ask Claude to relaunch the workflow.

The other way: ultracode for the whole session

The keyword runs one task. To have Claude plan a workflow for each substantive task in a session, type `/effort ultracode`. Turn it off with `/effort ultracode off`. Each request then costs more tokens and time, and you'll reach a session or weekly limit sooner. It also skips the large-run warning and, in auto mode, the first-launch approval.

Which jobs fit

The docs' examples are jobs with many similar pieces: checking every file for the same problem, migrating hundreds of files, and research where sources get cross-checked against each other. For a task that is one small change, a normal prompt is faster and cheaper.

Written on 17 November 2026 from Claude Code's own page on dynamic workflows (code.claude.com/docs/en/workflows) and its model configuration page. I haven't run these steps myself while writing this guide. Claude changes fast: if a menu has moved, look for the same words in the app.