

How to Save Tokens in Claude

HQ corporate guide for email, reports, policies, decks, and analysis

Master prompt

“

“I want to [task] to [success criteria]. Read the files I provided.
Ask clarifying questions before you start.
Keep the response concise and return only what I need.”



Why this matters

Claude often re-reads earlier context. Long chats, large files, and full rewrites cost more tokens.

| Less context + fewer turns + smaller outputs = **fewer tokens**

1



Start a new chat for a new workstream

Desired outcome

Avoid paying for irrelevant conversation history.

Action

If you switch from one task to another, open a fresh chat.

Example

Do not keep quarterly planning, policy review, and executive email drafting in one thread.

2



Ask for only the part you need

Desired outcome

Reduce output length and regeneration costs.

Action

Request one section, slide, table, or paragraph instead of a full redo.

Example

Rewrite the executive summary only. Keep the rest unchanged.

3



Batch related requests

Desired outcome

Reduce repeated context reloads.

Action

Combine related tasks into one prompt.

Example

Summarize this memo, list the decision points, and draft a short leadership update.

4



Send less source material

Desired outcome

Keep context small and relevant.

Action

Paste only the needed text and crop screenshots tightly.

Example

Instead of sending the full policy, paste only the section you need reviewed.

5



Edit or restart instead of stacking corrections

Desired outcome

Stop the conversation from growing unnecessarily.

Action

Revise the earlier prompt or restart from a clean point.

Example

Replace multiple follow-up corrections with one clean rewritten prompt.

6



End long chats at 20 messages

Desired outcome

Carry forward only what matters.

Action

Do not continue a chat beyond 20 messages. At 15–20 messages, ask Claude for a summary of key decisions, open questions, and next steps, then paste that summary into a new chat.

Example

Summarize this thread in 6 bullets: decisions, open items, next steps, and unresolved questions.

Do this every time >>

1



Define the exact work product

2



Send only the relevant context

3



Ask for the smallest useful response