

Context Compass

Task first, then model, then effort.

Part A context

What's in the context

Everything the model sees, re-sent and paid for on each request:

Tools & system prompt → **Persistent instructions** → **Memory** → **Files & tool results** → **Conversation history** → **Your message**

Chat, Cowork, Code:
where you use AI

Chat: Claude Chat / ChatGPT Chat — one answer.
Everyday agent: Cowork / Work — desktop folders & apps.
Code agent: Claude Code / Codex — repo & terminal.

**Instructions that
apply every time**

AGENTS.md / CLAUDE.md / project instructions: commands, conventions, vocabulary. No one-offs or secrets; >~200 lines is less reliable.

**Memory vs.
conversation history**

History: one chat, re-sent each turn. **Memory:** across chats; edit or delete. New topic? New chat. Sensitive? Temporary chat.

**More context isn't
better context**

Long documents **first**, your question **last**.

**Prepare the context
before you prompt**

Pick: only what the task needs. **Trim:** old versions, duplicates, unrelated threads. **Point:** file paths or links.

Part B models

Task → Model → Effort

How hard is the task? How often will it run? Is the default good enough?

Write 30 test cases from the spec → Claude Sonnet 5.5 / GPT-6.1 Sol → **medium**.

**How to read
a model name**

Claude Opus 5.5: family, tier, version.
GPT-6 Astra: family + generation, tier.
Tier matters more than version. Date = frozen snapshot.

Why every provider sells several models

Speed, cost and capability pull in different directions. Smallest → most capable:

Claude Haiku 4.5 → **Claude Sonnet 5.5** → **Claude Opus 5.5** → **Claude Fable 5.1**

GPT-6 Luna → **GPT-6.1 Sol** → **GPT-6 Astra**

Efficiency first: start small, test, upgrade only where it falls short.

Capability first: start strong, lower effort or downgrade where quality holds.

Effort: how hard the model works

Claude & OpenAI API levels; chat apps use thinking/reasoning settings.

none	Instant answers, no reasoning	OpenAI only
low	Simple tasks, sub-agents	Both
medium	Balance of speed and quality	Default: OpenAI, Opus 5.5
high	Complex reasoning and coding	Default: other Claude models
xhigh	Long agentic and coding runs	Both
max	Frontier problems only	Not every model

Start at the default; raise only if it falls short. Max can overthink and cost much more.

Myth Buster

The giant window → Paste what the task needs. The rest costs attention and money.

A prompt to keep

[Paste only the "Discount rules" section here: 2 pages, not all 40.]
Using the rules above, write test cases for stacked coupons.
If you need payment details, read docs/payments.md.