+64 27 459 8199
pranay@insyte-tech.com
www.insyte-tech.com

8 simple rules for using fewer AI tokens

An applied AI capability cheat sheet

Get useful answers without wasting your usage allowance - while protecting accuracy, privacy and accountability.

CLARITY SAVES MORE TOKENS THAN CLEVER WORDING.

Aligned with the Applied AI Capability Intelligence Framework

+64 27 459 8199
pranay@insyte-tech.com
www.insyte-tech.com

START HERE

Token efficiency is a work skill

Use less unnecessary input and output without weakening the decision.

What a token is

Tokens are the small pieces of information an AI assistant reads and produces. Your prompt, attached context, conversation history and the answer all contribute to usage.

YOUR INPUT	+	AI OUTPUT
Prompt + context	More relevant detail only	Answer produced

FRAMEWORK LENS

Efficient use is contextual

The right amount of context depends on the person, task, tools, environment and risk. A tiny prompt is not automatically a capable prompt. The useful target is enough context for an accurate, defensible result - with no extra baggage.

Use this five-part request pattern

1. OUTCOME	What useful result should this produce?
2. SOURCE	What information may the assistant use?
3. CONSTRAINTS	What boundaries, assumptions or risks matter?
4. CHECKS	How should important claims or artefacts be verified?
5. OUTPUT	What format and length do you actually need?

IMPORTANT

Do not optimise the wrong thing

Saving tokens is useful when it reduces noise, rework or unnecessary processing. It is not a reason to remove evidence, hide uncertainty, skip verification, expose sensitive information, or choose AI when a simpler deterministic method is more reliable.

+64 27 459 8199
pranay@insyte-tech.com
www.insyte-tech.com

RULES 1-2

Frame the job and control the context

RULE 1

Give the assistant a smaller job

Define the object, action and expected output. A smaller, explicit task reduces exploration and makes the result easier to judge.

CAPABILITY LINK: Task and outcome framing; AI interaction; acceptance judgement.

TRY THIS: Instead of "Help me improve my business", use: "Review this customer email and suggest three clearer responses."

DO NOT TRADE AWAY: Do not narrow the task so far that the business objective, audience or decision disappears.

RULE 2

Only provide relevant context

Before adding a document, ask whether it could change accuracy, constraints or the final recommendation. Include it only when the answer is yes.

CAPABILITY LINK: Context and information management; source selection; security and privacy.

TRY THIS: Provide the relevant section, name the source, and say why it matters. Remove duplicate or unrelated material.

DO NOT TRADE AWAY: Do not omit necessary evidence, provenance or risk information. Never paste sensitive data just because it is convenient.

OPERATOR CHECK

Can you defend the request?

A capable operator can explain why the task is suitable for AI, why each source is present, what a good answer looks like and what they will verify before using it.

+64 27 459 8199
pranay@insyte-tech.com
www.insyte-tech.com

RULES 3-4

Reset the conversation and bound the answer

RULE 3

Clean up long conversations

Old experiments, abandoned ideas and expired constraints travel forward in a long chat. Summarise the decisions, constraints and useful facts, then start fresh.

CAPABILITY LINK: Context management; workflow integration; evidence trails and handoffs.

TRY THIS: Create a short restart note with: objective, confirmed decisions, active constraints, source links and unresolved questions.

DO NOT TRADE AWAY: Do not lose decision history that a reviewer needs. Preserve important evidence outside the chat before resetting.

RULE 4

Set the size of the answer

State the number of options, maximum length, format and stopping point. This prevents useful work from being buried under output you did not request.

CAPABILITY LINK: Task framing; acceptance criteria; accountability for the final decision.

TRY THIS: Use boundaries such as: "Give three recommendations", "Keep it under 200 words", or "List decisions only; do not write the plan yet."

DO NOT TRADE AWAY: Do not use a word limit that prevents the assistant from stating material uncertainty, risk or supporting evidence.

QUICK RESTART TEMPLATE

Carry forward decisions, not clutter

Objective: ____. Confirmed decisions: ____. Active constraints: ____. Trusted sources: ____. Unresolved questions: ____. Required output: ____.

+64 27 459 8199
pranay@insyte-tech.com
www.insyte-tech.com

RULES 5-6

Stage the work and compress responsibly

RULE 5

Break large jobs into stages

Separate fact extraction, issue identification, outline review and final production. Review between stages so errors do not expand into unnecessary work.

CAPABILITY LINK: Work decomposition; verification; human oversight; failure handling.

TRY THIS: Use: 1) extract facts, 2) identify issues, 3) create an outline, 4) review the direction, 5) produce the final output.

DO NOT TRADE AWAY: Do not let staging become approval theatre. Review the evidence and direction at the points where a correction still matters.

RULE 6

Compress large inputs

Turn a long report, transcript or message history into a smaller information pack containing important facts, relevant examples, decisions and sections needing attention.

CAPABILITY LINK: Context management; provenance; source verification; risk-based checking.

TRY THIS: Keep source references beside each extracted fact. For consequential work, compare the compressed pack with the original before relying on it.

DO NOT TRADE AWAY: Compression can remove nuance. Retain exceptions, uncertainty, conflicting evidence and anything required for an accurate answer.

RISK RULE

Verification effort should rise with consequence

Low-risk formatting needs a lighter check than client advice, financial analysis, production code or a consequential business decision. Efficiency never removes human accountability.

+64 27 459 8199
pranay@insyte-tech.com
www.insyte-tech.com

RULES 7-8

Reuse good practice and choose proportionately

RULE 7

Reuse your instructions

Store stable guidance for audience, style, checks, guardrails and output format. Load the relevant instruction set instead of rebuilding it in every conversation.

CAPABILITY LINK: Builder capability; standardisation; guardrails; validation design.

TRY THIS: Create a reusable instruction set with five headings: Audience, Style, Checks, Guardrails and Format. Version it when the working rule changes.

DO NOT TRADE AWAY: Reusable instructions are not permanent truth. Review them when the tool, task, policy, audience or risk changes.

RULE 8

Match the setup to the task

Use lighter processing for classification, formatting, simple extraction and basic summaries. Reserve deeper processing for ambiguity, complex judgement and high-consequence synthesis.

CAPABILITY LINK: Tool and model selection; lifecycle judgement; risk changes the required standard.

TRY THIS: Ask: Could a rule, search, calculation or template do this more reliably? If AI is useful, choose the least complex setup that still meets the risk standard.

DO NOT TRADE AWAY: A lighter setup may reduce cost or protect a premium allowance, but it does not always reduce tokens processed. AI should not be the default.

BUILDER MOVE

Externalise judgement

When a request works repeatedly, capture not only the prompt but also its scope, source rules, acceptance criteria, validation steps, approval points, failure response and update owner.

+64 27 459 8199
pranay@insyte-tech.com
www.insyte-tech.com

REUSABLE CHECK

The two-minute token saver

Run this before sending an important request.

1	**Is the task specific?**	Name the object, action and result.
2	**Is every piece of context necessary?**	Keep what can change the answer; remove the rest.
3	**Is the conversation still clean?**	Restart with a decision summary when old context is interfering.
4	**Have I set the output boundary?**	Specify number, format, length and stopping point.
5	**Should I split the work into stages?**	Put a meaningful review before expensive final production.
6	**Can I reuse approved instructions?**	Load stable audience, guardrail, check and format guidance.
7	**Is this the right setup for the task and risk?**	Use a simpler method when it is more reliable.
8	**What evidence will make the result defensible?**	Define source checks, tests, acceptance and escalation.

From operator habit to builder practice

OPERATOR	BUILDER
Apply the eight rules to the current task. Check the output, manage risk and make the final decision.	Turn the rules into reusable templates, context packs, validation gates, permission boundaries, handoffs and recovery steps for repeated work.

USE RESPONSIBLY

This is a development aid, not a capability score

The checklist supports observable applied AI behaviours, but completing it does not prove universal capability. Capability remains contextual and should be judged from real work, evidence, risk-aware verification and responsible decisions.

Better token use starts with a clearer job - and ends with a result you can defend.