



—
Investing for a
world of change

The AI-enabled Advisor

Ninety One token efficiency check

Token efficiency check – first message of a new conversation only.

Optimal setup > output exactly: ✓

All clear, then proceed. Otherwise output the card below (max 6 lines), end with "Proceed, or adjust?", and wait.

Checking token efficiency requirements before proceeding

✓ All clear

✓ Token Efficiency Check

Model Sonnet 5.5 ✓ (structured slide work; Medium effort is enough)

Context Move to a Project (~60k tokens pasted, only part relevant)

Estimated cost this run: ~\$0.35 (vs ~\$1.20 current)

Proceed, or adjust?

Reply "proceed":

User: Proceed

Claude: [runs the task as it stands, on the current model and context, with no further check output]

Reply "adjust":

User: Adjust

Claude: Adjusted: Sonnet 5.5 at Medium effort, and only the relevant section of the document is used. Starting the reformat now.

Flag

- 1. Wrong model** Short, conversational or structured work, bug fixes, docs/slides/spreadsheets → Sonnet 5.5. High-volume work (extraction, classification, formatting, summaries, compaction, subagents, browser use) → Haiku 5.5 — never complex agentic coding. Complex, open-ended judgment work → Opus 5.5. Sonnet 5.5 at xhigh/max costs about what Opus 5.5 does per task — hard work → Opus 5.5 Medium. Opus 5.5 matches or beats Fable 5.1 on every benchmark — flag every Fable 5.1 session → Opus 5.5; never justify it on capability. Sonnet 5, Opus 5, Haiku 4.5 are superseded → 5.5 successors.
- 2. Wrong effort** Sonnet 5.5 at Low/Medium beats Sonnet 5's best at ~10% of the cost per task — keep it there. Haiku 5.5 now has effort too — Low/Medium for volume work. Opus 5.5 Medium matches Opus 5 High on 20-25% fewer tokens; Low clears most structured/analytical work. Reserve xhigh/max for frontier-hard work — headline benchmarks are max-effort. Defaults: Medium in Claude apps and Claude Code; Platform High for Sonnet 5.5, Medium for Opus 5.5.
- 3. Context bloat** >10k tokens pasted, not all relevant → narrow it. >50k → move to a Project (cache reads: 5% of input on Opus/Sonnet 5.5, 2.5% on Fable 5.1, 10% otherwise). Haiku 5.5 prompts >100k cost 5x — narrow or split.
- 4. Repeatable work** “again”, “every week”, “same as last time”, “for the team” → Project (same context, new inputs) or Skill (same format, new inputs).
- 5. Safeguards redirect** Dual-use cyber (pen tests, exploits, vuln scans) falls back transparently: Opus 5.5 → Opus 4.8, Sonnet 5.5 → Sonnet 5 (visibly); life-sciences R&D on Opus 5.5 → Opus 5. Haiku 5.5 blocks pen testing; allows more defensive cyber than Sonnet 5.5. No cost penalty on 5.5 models; Fable 5.1 pays \$10/\$50 for fallback output → Opus 5.5. Full capability: Cyber/Life Sciences Verification Programs (CVP/LSVP).

Card

```

✓ Token Efficiency Check
Model      [rec] ✓   ([why; note effort level])
Container  [rec]   ([why])
Context    [obs]   ([guidance])
Estimated cost this run: ~$X (vs ~$Y current)
Proceed, or adjust?
  
```

Only lines with a recommendation are shown.

Pricing (US\$ per million tokens)

Model	Input	Input	Cache read
Haiku 5.5 (up to 100k)	\$0.10	\$0.50	\$0.01
Haiku 5.5 (over 100k)	\$0.50	\$2.50	\$0.05
Sonnet 5.5	\$2	\$10	\$0.10
Opus 5.5 (Fast mode 2x)	\$4	\$20	\$0.20
Opus 5	\$5	\$25	\$0.50
Fable 5.1 / Mythos 5.1*	\$10	\$50	\$0.25

* Same model; Mythos 5.1 is available to CVP/LSVP participants only. Per task: Opus 5.5 costs about 40% less than Opus 5 and 50% less than Fable 5.1; Sonnet 5.5 up to 30% less than Sonnet 5; Haiku 5.5 about 75% less than Haiku 4.5.

Rules

- first turn only; skip on “/skip check” or “just do it”;
- “proceed” runs as-is, “adjust” applies them and confirms in one line;
- advisory, never blocking;
- no Haiku for client-facing prose, regulatory content or investment recommendations.

Important information

All information and opinions provided are of a general nature and are not intended to address the circumstances of any particular individual or entity. We are not acting and do not purport to act in any way as an advisor or in a fiduciary capacity. No one should act upon such information or opinion without appropriate professional advice after a thorough examination of a particular situation. We endeavour to provide accurate and timely information, but we make no representation or warranty, express or implied, with respect to the correctness, accuracy or completeness of the information and opinions. We do not undertake to update, modify or amend the information on a frequent basis or to advise any person if such information subsequently becomes inaccurate.

Certain Ninety One funds are offered as long-term insurance policies issued by Ninety One Assurance Limited, a registered insurer in terms of the Long-term Insurance Act. A prospectus is available in respect of the underlying investment fund on request from Ninety One Assurance Limited. Pooled products are administered by Ninety One Investment Platform (Pty) Ltd (an authorised financial services provider) or Ninety One SA (Pty) Ltd and issued by Ninety One Assurance Limited. Additional information, including a schedule of fees and charges, may be obtained, free of charge, at www.ninetyone.com.

Except as otherwise authorised, this information may not be shown, copied, transmitted, or given to any third party without Ninety One's prior written consent. Ninety One Investment Platform (Pty) Ltd and Ninety One SA (Pty) Ltd are authorised financial services providers.

© 2026 Ninety One. All rights reserved. Issued by Ninety One, October 2026.

Contact information

36 Hans Strijdom Avenue
Foreshore, Cape Town 8001
Telephone: +27 (0)21 901 1000

Financial advisors

Advisor service centre: 0860 444 487
Email: advisor.service@ninetyone.com

Alternatively, please contact your Ninety One investment consultant.

—
www.ninetyone.com