Cut your AI costs 75%. Keep the same quality.

Haakon lets you run smaller, faster models — like Haiku and Gemini Flash — instead of expensive frontier models, and still match or beat their performance. Your AI spend drops to a fraction of what it was.

BY THE NUMBERS

75%
Lower AI spend
vs. using one expensive model for everything

3×
Faster task completion
On complex, multi-step requests

4×
More done per AI budget
Same monthly limit, 4× the work done

94%
As good as the expensive model
Vs. top models from OpenAI and Anthropic

HOW IT WORKS

Three things Haakon does for you

01

Unlocks more from every model

Cognitive lenses plug structured reasoning into any model. Use cheap models like Haiku or Gemini Flash for frontier-level output, or push frontier models even further.

02

Fits into tools you already use

It works inside Claude, ChatGPT, Gemini, Cursor, and VS Code. No new app for your team to install or learn.

03

Runs wherever you need it to

Use it in the cloud for speed, or run it entirely on your own servers with nothing sent outside your walls.

FROM THE RESEARCH

What the data shows

→ **Your AI provider can quietly downgrade you**
When providers hit capacity, teams on a single model saw quality drop — with no warning and no way to fix it.

→ **The right framework beats a bigger model**
Applying a cognitive lens to a fast, cheap model matched the quality of defaulting to the priciest model every time.

→ **People actually use tools they already know**
Teams adopted AI far faster when it showed up inside apps they already had open, not a new login to learn.

→ **Locking data in-house barely costs you anything**
Fully private, offline deployments for government and defense teams performed within 8% of the cloud version.

SETUP

Plugs in, doesn't bolt on: Haakon connects to the AI tools you already have through an open standard (MCP), so there's no custom integration project to run.