FREE ONE-PAGER · FROM THE FILM COMPUTE VS MEMORY

The Founder's Guide to AI Costs

The cost of intelligence keeps falling, from the data centre down to the laptop on your desk. Build your business for where that cost is going, not for where it is today.

01 · WHY AI COSTS WHAT IT DOES

The chip is waiting for memory

Chat models write one token at a time and re-read the whole model for each one. Long prompts and many users fill memory with running notes (the KV cache).

Compute growth 1996–2023 — ≈ 60,000×



Memory bandwidth growth — ≈ 100×



Log scale. Gholami et al., "AI and Memory Wall", IEEE Micro 2024.

02 · WHERE THE COST IS GOING

Down, and onto smaller machines

- **Parallel writing.** Stanford's Next-Concept Modeling ran 10–31× faster than same-size diffusion models. Research only, tested at 100M–725M parameters.
- **Smaller machines.** Capable models already run on an ordinary laptop with the internet switched off.
- **More use.** When answers get cheaper, people ask for more of them. Plan for volume, not just price.

Kumbong et al., arXiv 2610.14162 (Oct 2026).

03 · EIGHT MOVES

1. **Price for the curve**
   Assume each answer gets cheaper every year. Let margins widen as it falls. Avoid long commitments at today's prices.
2. **Smallest model first**
   Give each task the smallest model that does it well. Send only the hard cases to the biggest one.
3. **Keep context lean**
   Long context fills memory and costs on every call. Send what the task needs, not the whole history.
4. **Cache what repeats**
   System prompts, documents and common answers that repeat should be reused, not regenerated.
5. **Batch what can wait**
   Reports, tagging and summaries rarely need an answer in a second. Queue them and run them together.
6. **Run local where it pays**
   Private data, offline work and high volume are good reasons to run a model on your own machine.
7. **Measure cost per task**
   Track what one finished task costs, retries included. Tokens are an input. Tasks are the product.
8. **Review every quarter**
   Prices and models change fast. Re-test your routing and your vendor every three months.

04 · YOUR COST-PER-TASK WORKSHEET

```
cost per task = (tokens in × price in) + (tokens out × price out) + retries
```

TASK	MODEL	TOKENS IN	TOKENS OUT	COST / TASK	TASKS / MONTH	MONTHLY COST
e.g. reply to a lead						

Watch the film and read the essay: nikhilsharma.com/soch/vimarsh/compute-vs-memory
Building on falling AI costs? **NS Transform 2027** · nikhilsharma.com/transform

For education. Not investment advice. © Nikhil Sharma, November 2026.

NIKHIL SHARMA
@NIKHILSHARMAHQ