Corta: Benchmark-trained router for AI

One-liner

Corta sits as a decision-making layer between your app and the universe of models. It automatically directs any query to the best model across local or cloud, optimizing for cost, latency and performance, with just one line of code.

Why is using one big model for everything such a bad idea?

Let's start from first principles.

1. Sending every task to the same model isn't efficient

On the surface, yes. One giant general model can handle everything (email summaries, legal drafting, code generation). But in practice it's a blunt tool.

Running all tasks through GPT-5 is like plugging your lightbulb, Tesla, and factory machinery into the same power source. It technically works, but it's not the right fit for most jobs. True efficiency in AI comes from using the right model for the right task.

2. Using GPT-5 for everything is the wrong approach

Running everything through a giant general model is expensive and inefficient. Top-tier models often cost $0.10 to $0.60 per query because they require long prompts, large context windows, and GPU-heavy infrastructure. But price isn't the only issue.

Benchmarks show that smaller, specialized models consistently beat GPT-5 in domains like math, medical, legal, and code at 10 to 100 times lower cost.

So you're not just overpaying. You're also underperforming. It's like hiring a world-class brain surgeon to take your temperature: technically possible, financially reckless, and usually less effective than the specialist tools built for the job.

3. Better models exist, but adoption lags behind

In 2021 there were a handful of major models. By 2025 there will be over 10,000. Some are 10x faster, some 100x cheaper, and many beat GPT-5 in narrow domains. Yet most teams stick with one expensive model or hack together fragile routing logic that doesn't scale.

AI capability itself has plateaued. As Andrej Karpathy put it, "If you are prompting a model a lot, it's usually the wrong model." The future is not about squeezing more out of a single general model. It's about picking the right specialist for the job, the same way you'd see a cardiologist instead of a general physician if you had a heart problem.

The problem is that no one has the infrastructure to do this at scale.

So what's the real solution here?

The answer isn't picking one giant model or stitching together brittle routing rules.

The real solution is intelligence. Corta sits as a decision-making layer between your app and the universe of models. Every query is analyzed in real time, and the router selects the model that best balances cost, latency, accuracy, and domain fit. It learns continuously through benchmarks and feedback, so routing gets smarter with every request.

This isn't another infra tool. It's not logging, retries, or caching. This is a new layer of intelligent decision-making, the routing brain for AI workloads.

Developers just write:

```
corta.route("query")
```

That's it. One line. Corta handles everything under the hood.

Why is this the right moment to build Corta?

AI is shifting from the capability curve to the efficiency curve. Bigger models aren't translating into better outcomes anymore — they're translating into higher costs, longer prompts, and diminishing returns. As Andrej Karpathy put it: *"If you are prompting a model a lot, it's usually the wrong model."*

1. **Bigger models used to lead to better outputs**

   Not anymore. Research from NVIDIA and Stanford shows that smaller, task-tuned models now outperform general giants like GPT-4 and GPT-5 in math, medical, and code benchmarks, often at 10–100x lower cost. As Nathan Lambert (Anthropic) noted on Twitter, "Specialist models are starting to quietly crush generalists." Bigger is no longer better.

2. **Cost per query matters more than ever**

AI startups and enterprises alike are bleeding millions each year on model usage. OpenAI's GPT-4 Turbo runs up to $0.60 per query in production settings. Multiply that by billions of queries, and you have infra spend that dwarfs revenue. Efficiency isn't optional; it's survival. As Emad Mostaque said, "The real AI race is the cost race."

3. Choosing the right model is no longer humanly possible

In 2021, there were a handful of frontier models. In 2025, there will be over 10,000. Some are faster, some are cheaper, some win in narrow domains. But no human team can benchmark, evaluate, and switch between them at runtime. Choosing the right model has become a computation problem, not a manual one.

4. AI is moving to the edge

Workloads are shifting from cloud to devices — phones, wearables, robots. Large models drain batteries and require network calls. Local models run fast and cheap, but aren't always enough. What's missing is a brain that knows when to run locally and when to escalate. As Chris Lattner (Modular) tweeted, "The future isn't more GPUs, it's smarter scheduling."

The moment is clear: AI capability has plateaued, efficiency is the next frontier, and intelligent routing will decide who survives the cost curve. Corta is building that layer.

How is Corta different from other "router" tools?

Most other routing attempts are shallow wrappers around existing APIs that add caching, retries, or logging.

Corta is different. We optimize for **real task performance**, using real benchmarks.

We've built:

1. A **classifier** that analyzes each query and predicts domain, complexity, and required accuracy
2. A **router** that selects the best model based on:
3. Cost per token
4. Latency
5. Benchmark performance (from LM Arena, HELM, Hugging Face, etc.)
6. Contextual and domain fit
7. A **feedback loop** that gets smarter with every routed query

This is not infra. It's intelligence.

We're building the real-time performance layer that sits between the user and the AI model like a Stripe or Cloudflare for model selection.

Aren't there already players in this space? What are they getting wrong?

There's excitement in the space, but no one is solving the actual routing intelligence problem.

1. **OpenRouter** lets users pick models, but the routing is manual and static.
2. **Portkey, Martian, Humanloop, Baseten** offer infra tooling, not dynamic, benchmark-aware routing.
3. **LangChain and LlamaIndex** are developer frameworks, not routing engines.

The missing piece in the AI stack is a **performance-optimized router** that selects the best model automatically. That's exactly what Corta is building.

Who actually needs this and how big is the opportunity?

Corta is a horizontal layer that will sit in every AI product, across every vertical. Here's how we break it down.

1. Device Manufacturers (Phones, Robots, IoT)

They need to balance power consumption, latency, and performance between local and cloud inference. Corta's SDK makes that seamless.

> Example: An AI wearable company can use Corta to route simple voice commands to a local model, and escalate complex queries to the cloud only when needed.

2. AI-Native Startups (Legal, Code, Health, Finance)

These teams care deeply about unit economics. Saving $0.01 per query can mean millions per year.

> Example: A Legal AI startup uses Corta to route standard clauses to Claude Sonnet, and only uses GPT-4 Turbo for complex litigation language.

3. Cloud & Model Providers

They want to support more models, but struggle to optimize traffic between them. Corta provides white-labeled routing logic they can embed in their stack.

4. Developers

Over 30 million developers globally. More than 5 million are already experimenting with AI. They want simple tools that work.

corta.route(query)

That's the Stripe motion - bottom-up, dev-first, easy to adopt and scale.

How does Corta make money?

Corta makes money every time someone routes a query. Clean, scalable, and usage-based.

1. API / SDK Usage

1. Stripe-style pricing
2. $0.0001 to $0.01 per routed query
3. Free tier for hobbyists, Pro and Enterprise tiers for scale

2. Enterprise Licensing

1. On-prem deployment
2. Custom benchmarks
3. Dashboards for latency, cost, audit logs, and performance
4. Priced between $10K to $500K+ per year

3. Marketplace Fees

1. Long-term, Corta can charge a % fee for routing traffic to partner models like Mistral, Anthropic, or Cohere
2. Think of it as a neutral Stripe for model providers

4. Edge SDK Licensing (Future)

1. Offer routing SDKs to device manufacturers
2. Run local models by default and escalate only when needed
3. Think AI-aware OS for on-device inference

What's the path to making this a critical part of the AI stack?

We're building Corta in three phases, each compounding into defensibility:

Phase 1: Smart Router
Launch an API and SDK that developers can drop in with one line of code. Queries are routed

using public benchmarks and provider metrics, giving immediate savings on cost and latency. We start charging from day one.

Phase 2: Feedback Flywheel
Every routed query creates data: input → model → outcome. By learning from user feedback and real-world success rates, we build a proprietary dataset no one else has. Routing gets smarter automatically, creating a performance moat.

Phase 3: Benchmark Standard
As volume grows, Corta becomes the de facto evaluation layer. We'll launch real-time benchmarks across domains and incentivize developers to contribute outcomes. Corta evolved into the standard performance metric for AI routing, the way Stripe standardized payments and Cloudflare standardized delivery.

At that point, Corta isn't just infra. It's the intelligence layer every AI product depends on, sticky, defensible, and impossible to build around.

Who's behind this and why are they the right people to build it?

Juan Blanco, Co-Founder

1. Led Data and ML projects at AWS for enterprise clients like Santander and Zara.
2. Founded Saiki (AI + psychology startup, MassChallenge Finalist 2019).
3. Built and ran the data function at dOrg (100+ clients, $11M revenue).
4. Closed $20M+ in B2B deals across AI and Web3.

Faraaz Baig, Co-Founder

1. Raised over $11M for a drone startup at 20.
2. Emergent Ventures Fellow (Tyler Cowen) & Cerebras Fellow.
3. Built and programmed world's first blended wing body tailsitter which was later patented.
4. Built viral AI tools with over 250K+ organic impressions and 5K+ users.
5. Founded Spill, a minimalist freewriting app that ranks 2nd on product hunt.

Ahmed Baig, Co-Founder

1. Previous UCI engineering ML and data analysis researcher.
2. Developed ML classification models for REACH, a smart wearable device. (Stella Zhang New Venture Finalist, Beall Applied Innovation Winner).
3. Built QA and cloud automation web apps for AVEVA; still in use today.

4. Generated 1.5M+ views across social media facilitating $30k+ in sales within 2 months.

What's the big picture? Why does this matter now more than ever?

In 1997, the internet needed DNS, a system to intelligently route traffic across a chaotic, growing network of servers. In 2025, AI needs the same thing.

We are no longer in the age of one model to rule them all. We are in the age of many, and soon millions, as models proliferate, specialize, and move to the edge. The bottleneck is no longer generation. It is orchestration. The AI capability curve has plateaued. The next curve is efficiency, and it will be defined by who can route intelligently.

Corta is building the routing brain for this world. With a single line of code, every developer, product team, and enterprise can access the smartest, fastest, and most cost-effective model automatically.

This is not logging, retries, or caching. It is the missing layer of intelligence in the AI stack. And once it is in, it is everywhere. As fundamental as Stripe for payments, Cloudflare for delivery, or DNS for the web.

Now is the time to build it.