HIETANEN CONSULTANCY · CUSTOMER BRIEFING · SEPTEMBER 2026



Fuuga

The sovereign LLM stack for .NET.

Train, fine-tune, compress and serve your own language models on .NET — no Python, and no data leaving the building.

hietanen.co.uk · nuget.org/packages/Fuuga

THE PROBLEM

Enterprises rent their AI, and cannot maintain the AI they build.

Rented models, no control

Data leaves the company, every request is billed, and the vendor can change or retire the model under a live product.

In-house models today are a specialist's side project

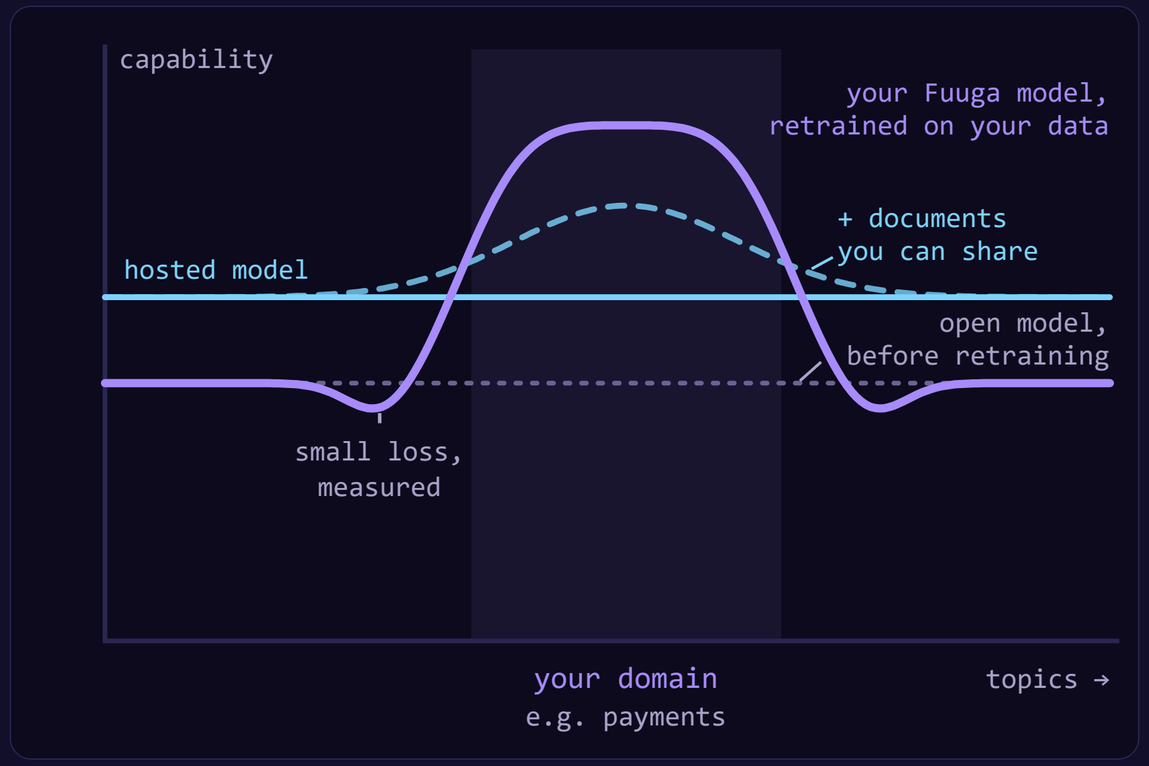
A data scientist builds them in Python tools the engineering team does not use. When that person leaves, nobody can maintain them.

Regulation asks who owns the model

For a high-risk use such as credit scoring, the EU AI Act requires a technical file describing the training data. DORA requires an exit plan for each critical ICT supplier.

THE IDEA

Ahead where your business is.



Retrained on your work

Your terms, decisions and code conventions become its defaults.

Grounded in all your documents

Searched in-house, including the ones that may never leave.

Measured both ways

The gain and the small loss, on your own held-out tasks, before release.

THE PRODUCT

Your own AI model, built and maintained by the engineers you already have.

The complete toolchain to make a language model, or retrain an existing one — on .NET, the platform banks and insurers already run.

- Starts from open models you know — Llama, Qwen, Gemma, Mistral, Phi, DeepSeek, Kimi — and teaches them your data
- Runs on ordinary GPUs or CPU servers you already own
- Speaks the OpenAI and Anthropic interfaces; nothing is rewritten

for the technical reader — the common path

```
> fuuga scaffold finetune
> fuuga transfer --recipe my.json
> fuuga sft --recipe my.json
> fuuga eval --recipe my.json
> fuuga export gguf --tokenizer tok/
> fuuga-serve --checkpoint best/
```

HOW IT WORKS

Start from an open model. Make it yours. Keep it in-house.

01
Ingest
Company data becomes a traceable corpus.

02
Transfer
Start from nine open model families.

03
Fine-tune
Teach it the company's work on ordinary GPUs.

04
Compress
Shrink it to run on servers already owned.

05
Serve
OpenAI- and Anthropic-compatible, in-house.

Every step is measured, so an agent can run the loop — and only keeps a model that is better.

WHAT RUNS TODAY

Not a roadmap. A working stack in the repository.

Area	In the codebase now	Status
Training	Modern transformer architectures, models larger than the GPU	Shipped
Fine-tuning	SFT, DPO, GRPO, LoRA and QLoRA, typed judge	Shipped
Model operations	Transfer, distillation, merging, quantisation, export	Shipped
Serving	OpenAI and Anthropic APIs, batching, guardrails, MCP	Shipped
Metered hosting	Per-token pricing and prepaid credit per API key	Shipped
Agents	Sandboxed tools, cost-aware routing, knowledge-graph RAG	Shipped
Quality	Unit and property tests across every module	Shipped

WHY NOW

Open models caught up and regulation arrived — yet Microsoft's .NET path is still a rented one.

Open models are good enough

Adapted to a company's own cases, open weights beat a generic hosted model.

Hosted prices must rise

Today's per-token prices are subsidised by investor capital.

Sovereignty is required

Credit scoring becomes high-risk under the EU AI Act in December 2027; DORA has applied to finance since January 2025.

.NET's path is rented

Microsoft's route is Copilot and Azure OpenAI: OpenAI and Anthropic models, hosted, not owned.

WHERE IT PAYS FIRST

Typed decisions: the part of AI a business can actually measure.

The data they have

Outcomes, not a corpus

Every company records outcomes — approved, declined, fraud. Those are the training labels.

The number they get

An answer to "is it better?"

A decision has ground truth, so accuracy and calibration are measured on the customer's own cases.

The line it moves

Act, or escalate

Confident cases are handled by software; the rest go to a person. The share automated is the return.

Nothing here generates text. One pass returns a calibrated probability for each option — logged, audited and priced.

A PUBLIC SHOWCASE

The pipeline works end to end, and every result is measured.

Decisions

59% → 90%

Accuracy on unseen cases

Open Laya, retrained in Fuuga in 6.5 minutes. The hosted, closed Jev on the same cases: 58.5%.

Acts on its own

66% → 94%

Right when it decides alone

It now decides 92% of cases without a person.

Measured both ways

0 of 6

Public benchmarks fell

A small donor on one 12 GB card. Larger Gemma and Phi follow.

Public datasets, repeatable from the repository. The retrained model is available for a live demo.

ADOPTION PATH

The bill is the business case. Move a workload only when the numbers say so.

01

Connect

Same OpenAI and Anthropic interfaces; point the app at a new address.

02

Route

Sensitive and routine requests in-house, the rest to the subscription.

03

Learn

Train the in-house model on the company's own traffic.

04

Switch

When the measured comparison passes, the subscription shrinks.

A rising per-request bill becomes a fixed cost — and the weights leave in standard formats, so it runs anywhere, including without us.

`Hietanen Consultancy · Fuuga · 10 · fuuga-serve --api-compat both; cost-aware router; distill --teacher-endpoint; export gguf / onnx`

COMPETITION

Today's options are a Python team or a rented model. Fuuga adds a third.

What the buyer gets	Fuuga	Python stack	.NET libraries	Copilot and hosted APIs
Train and fine-tune a language model	Yes	Yes	Small classifiers	On their servers
Data stays on your hardware	Yes	Yes	Inference only	No
Maintained by the .NET team	Yes	New team	Yes	Yes
Training provenance	Built in	Assemble it	—	Vendor statement
You own the weights	Yes	Yes	Import only	No
Cost as usage grows	Fixed	Fixed	Fixed	Per request

Python: PyTorch, HuggingFace, vLLM · .NET libraries: Semantic Kernel, LLamaSharp, ML.NET · Hosted: Copilot, Azure OpenAI, Mistral

TEAM AND METHOD

One engineer, AI-assisted, shipped the stack. The company grows as customers sign.



Tuomas Hietanen — Founder and CEO

Nine years as CTO of two regulated lenders, UK and US. Maintainer of SQLProvider and LINQKit.

Who owns what	Today	Next hire
Platform	Founder, AI-assisted	Engineers
ML delivery	Founder and network	ML engineers
Enterprise sales	Founder	Head of sales
Compliance	Founder	Compliance lead

Identified, and named once agreements are in writing: a second senior team member, and partners for data, compliance and ML advisory. The network is senior engineers we have shipped with, joining on contract. The founder hands over in the order shown, platform first, as signed revenue pays for each hire; public code, full tests and written recipes let them run the loop from the first day.

WORKING WITH US

Start with a pilot on outcomes you already record.

01 · Pilot

Fixed scope

A calibrated decision engine in production on your own outcomes.

02 · We run it

In your environment

Retraining, drift monitoring, evidence for your supervisor, on-call.

03 · You run it

Supported builds

Signed releases, an SLA, provenance — and onboarding each new engineer.

04 · Leave any time

Nothing locked in

Public-domain code, standard weight formats, every recipe written down.

Tuomas Hietanen · hello@hietanen.co.uk · hietanen.co.uk/projects/fuuga/

Speaker notes · Fuuga Customer Briefing

01 · Fuuga

Every enterprise wants a model of its own, and almost none on .NET can build one today. Fuuga is the complete pipeline that makes it possible, on the platform their engineers already run.

02 · Enterprises rent their AI, and cannot maintain the AI they build.

Companies rent their AI, and the AI they build themselves is a specialist's side project nobody can maintain. Regulation now asks who owns the model.

DETAIL

— Rented: records leave the company, every request is billed, and the vendor can change or retire the model.

— Built in-house: a separate Python toolchain, shipped once, orphaned when its author leaves.

— Regulation: credit scoring needs an EU AI Act technical file on the training data; DORA needs an exit plan from every critical ICT supplier.

ANSWERS

— A hosted vendor's compliance statement? A statement, not the training data the technical file describes.

03 · Ahead where your business is.

An open model, retrained on the company's own work and grounded in all its documents, gets ahead of a hosted model where the business is. The cost, a little general ability just outside the domain, is measured before release.

DETAIL

— Retraining teaches skills and language: the company's own terms, decisions, message formats and code conventions become the model's defaults, not the industry average. Facts that change are looked up in the company's documents when the question is asked.

— The dashed line is a hosted model given the documents allowed to leave the network; the in-house model searches all of them.

— Adapters (LoRA) and general examples in the training mix keep the loss small; fuuga eval --compare reports both numbers.

— Even where the curves are level, the data stays inside, the cost is flat and no vendor can retire the model.

— The curves are an illustration; each pilot measures them on the customer's own tasks.

ANSWERS

— What gets worse? A little general ability just outside the domain, reported with the gains. In the public showcase, none of six benchmarks fell.

— Why not give a hosted model our documents? It helps with facts, not with how the company decides, and not for classified or regulated data.

— Questions outside the domain? The customer's choice: the same model, a hosted one, or both.

04 · Your own AI model, built and maintained by the engineers you already have.

Fuuga is the whole toolchain to make a language model or retrain an existing one, on .NET, maintained by the engineers the company already employs.

DETAIL

— Starts from open models: Llama 3, Qwen 2 and 3, Gemma 3 and 4, Mistral, Phi-3 and Phi-4, DeepSeek-V3 and Kimi K2, with GLM and Kimi K3 experimental. GGUF files load directly; ModernBERT is the base for typed decisions.

— Speaks the OpenAI and Anthropic interfaces, so existing applications connect unchanged.

— Weights export to the standard formats, so nothing is trapped.

— No Python means no second toolchain owned by a specialist.

05 · Start from an open model. Make it yours. Keep it in-house.

Five steps, left to right: ingest, transfer, fine-tune, compress, serve. Every step is measured, so an agent can run the loop and keep only a model that is better.

DETAIL

— Ingest: a hashed corpus; every checkpoint records which sources trained it.

— Transfer: nine model families (Llama, Qwen, Gemma, Mistral, Phi, DeepSeek with Kimi K2, GLM and Kimi K3 experimental, ModernBERT), also from GGUF files, or a small model from scratch.

— Fine-tune: LoRA and QLoRA on consumer GPUs, DPO, GRPO.

— Compress: distil, merge, prune, quantise; export GGUF or ONNX.

— Serve: OpenAI- and Anthropic-compatible APIs with batching, guardrails, tool calling, MCP and A2A.

06 · Not a roadmap. A working stack in the repository.

Everything on this slide is in the repository today, under test, and can be walked through live.

DETAIL

— Training: RoPE, GQA, MLA and MoE; AdamW and Muon; CPU and NVMe offload for models larger than the GPU.

— Fine-tuning: SFT, DPO, GRPO, LoRA and QLoRA, and a typed judge for old-versus-new comparisons.

— Metered hosting: per-token prices and prepaid credit per API key, with the price list in OpenRouter's format.

— Also shipped: Stable Diffusion 1.5 from scratch, text-to-video, OpenTelemetry, bit-exact export round-trips.

07 · Open models caught up and regulation arrived — yet Microsoft's .NET path is still a rented one.

Open models are good enough, hosted prices have to rise, regulation asks for sovereignty, and Microsoft's path for .NET teams is still a rented one.

DETAIL

— Open weights from Llama, Qwen, Gemma, Mistral, DeepSeek and Kimi, adapted to a company's own cases, beat a generic hosted model at its tasks.

— Hosted inference is priced below cost with investor money; analysts estimate prices would need to rise several-fold.

— Copilot and Azure OpenAI run OpenAI and Anthropic models, rented per request; none gives a company a model it owns.

— QLoRA and offloading adapt a useful model on one workstation GPU.

08 · Typed decisions: the part of AI a business can actually measure.

Typed decisions answer the two objections every buyer raises: we have no training data, and we cannot tell whether a new model is better. Recorded outcomes are the training data, and a decision has a right answer to score against.

DETAIL

— Outcomes: approved, declined, fraud, which team took the case.

— Scored with accuracy, log loss, Brier score and calibration error on held-out cases.

— Above the confidence threshold the software acts; below it, a person does. The share automated is the return.

— For prose, the judge is typed too: old against new as one closed choice, asked both ways round.

ANSWERS

— How do you know the retrain helped? The typed judge compares old and new on held-out cases; public benchmarks are checked as well.

— What accuracy do customers get? Measured per customer, on their own held-out cases, inside their environment.

09 · The pipeline works end to end, and every result is measured.

A public showcase: the pipeline works end to end, and every result is measured. Retraining the open decision model Laya took accuracy on unseen emotion cases from 59% to 90%, far ahead of the hosted, closed Jev at 58.5%, and none of six public benchmarks fell on the retrained text model.

DETAIL

— Laya: 58.7% as published, 89.9% after 6.5 minutes on 4,000 rows on one RTX 4070, exported back in its own format.

— Jev, the same 2,000 questions through its API: 58.5%. Retrained Laya is right on 684 of the 741 cases they decide differently ($p < 0.0001$).

— Without a person, when at least 90% sure: as published, it decided 68% of cases alone and was right on 66% of those; retrained, it decides 92% alone and is right on 94%.

— Qwen3-1.7B, 80 MB of private text, 2 h 18 min: five benchmarks held or rose; TruthfulQA rose eight points.

ANSWERS

— Was Jev treated fairly? Same questions and option names; its documentation recommends a description per option, which may raise its score.

— Why 1.7B? To prove the path on one 12 GB card. Larger Gemma and Phi donors need more time or a bigger GPU; Fuuga offloads GPU memory to RAM and SSD.

— Does the text model beat OpenAI or Anthropic? It is compared with itself; that comparison runs in each pilot, on data no hosted model has seen.

10 · The bill is the business case. Move a workload only when the numbers say so.

An existing subscription is not replaced on day one. Connect, route, learn from the incumbent, and move a workload only when the measured comparison passes.

DETAIL

— Connect: the OpenAI and Anthropic interfaces are served together, so applications only change address.

— Route: the cost-aware router sends sensitive or routine requests in-house and the rest to the subscription, within a budget.

— Learn: endpoint distillation trains the in-house model on the hosted model's answers to the company's own prompts.

— Switch: a typed judge on the company's own work decides; the weights leave as GGUF, safetensors or ONNX.

11 · Today's options are a Python team or a rented model. Fuuga adds a third.

Today's options are a Python team or a rented model. Fuuga is a third: train, own and run the model on .NET, with the team the company already has.

DETAIL

— Python stack (Unsloth, Axolotl, HuggingFace): excellent, and a second toolchain with a specialist team.

— .NET libraries (Semantic Kernel, LLamaSharp, ML.NET) call models or fine-tune small classifiers; TorchSharp is the tensor library Fuuga is built on.

— Copilot and Azure OpenAI rent models per request; Fuuga implements Microsoft.Extensions.AI and sits alongside them.

— Hosted typed-output engines such as Jev validate the category and keep the weights. In the public showcase, Laya retrained in Fuuga scored 89.9% against Jev's 58.5% on the same 2,000 cases.

ANSWERS

— Over a hosted API? The records stay inside, the cost is fixed, and no vendor can retire the model.

— Over a Python stack? One toolchain and one team, with provenance by default.

— Over building on TorchSharp? Everything above the tensor library is the work: import, fine-tune, judge, compress, export and serve, all under test.

12 · One engineer, AI-assisted, shipped the stack. The company grows as customers sign.

One founder directing AI coding agents built and tested the stack, with a network of senior engineers for delivery. The code, the tests and the history are public, so the claim can be checked.

DETAIL

— Tuomas Hietanen: nine years CTO and co-founder of Finular Group, running two regulated lenders, Fund Ourselves in the UK and WeLendUs in the US; lead engineer at Basware; maintainer of SQLProvider and LINQKit; published .NET author.

— Funded by consultancy income and the founder's own time.

— Roles pass to hires as signed revenue pays for them: platform engineering first, then ML delivery and sales, then compliance.

— Key-person risk is covered by public, tested code and rerunnable recipes.

ANSWERS

— Who else? A second senior team member and partners for data, compliance and ML advisory are identified; names go on the slide once agreements are signed.

— Can it be verified? The code, tests and commit history are open; a technical partner can check it in an afternoon.

13 · Start with a pilot on outcomes you already record.

Start with a pilot on a decision you already record outcomes for. It ends with a calibrated engine in production, and you can leave at any time with everything.

DETAIL

— Pilot: fixed scope, trained and measured inside your environment; what leaves is the recipe, never the records.

— We run it: retraining, drift monitoring, the evidence pack your supervisor asks for, and someone on call.

— You run it: signed builds, long-term support, advisories, an SLA, and onboarding when people change.

— Leaving: public-domain code, standard weight formats and rerunnable recipes, which is also the exit plan DORA asks for.

ANSWERS

— What do you need from us? A decision with recorded outcomes, an environment to train in, and one process owner.

— Is this one person? One engineer full time, with senior engineers from the network when the work needs them.

— Can a partner deliver it? Yes: certified implementers deliver it and escalate to us.