Three Models. Three People.

An AI framework for wealth managers and entrepreneurs, helping you gain efficiency and preserve autonomy.

Dr. Andreas K. Janoschek · Founder & CEO, Gerevest.AI, Geneva · 1 September 2026

IN BRIEF

- Independent Swiss wealth managers are using AI. Yet a study[1] commissioned by Innosuisse shows that as of early August 2026, a good two thirds of all applications are still stuck in experimentation mode. Almost a quarter are pilot projects, but only about 10% are scaling.
- **This paper presents a framework that takes any AI project from planning through build and into operations.**
- Frontier models are too expensive for continuous operations, and your sensitive data travels through data centers.
- Open source models can be configured by you and run on your local hardware in full compliance with data protection. But these models are not as smart as frontier models and only suited to limited tasks.
- **We resolve this contradiction. We show the optimal model class per use case and name realistic hardware options for self-hosting.**
- Also since August 2026, Article 50 of the EU AI Act applies. We cover the impact of regulation (transparency duties, AI literacy training) and show where you find Finma and the EU AI Act in our framework.

1 Broad interest, many pilots, few scaled operations

You know the feeling. The demo runs, the team is excited, the board nods. Three months later the same employee copies the same reporting by hand from the same three systems. The demo is still alive, on the laptop of the intern who built it. The intern has been back at university since September.

This state of affairs now has numbers. On August 3, a study by the Lucerne University of Applied Sciences and Arts, together with the Zurich wealthtech Finup, appeared, commissioned by Innosuisse. 49 percent of the total of 41 Swiss wealth managers surveyed between April and June 2026 already use AI in selected processes, another 39 percent are in the exploration or planning phase. Only 12 percent have merely discussed the topic so far, without pursuing concrete initiatives.

Up to this point the numbers sound promising. But of the reported AI applications, 66% are merely experiments with public tools such as ChatGPT, Copilot or Claude. 23% are pilots. 11% run at scale. 83% call AI a strategic opportunity.

Read these numbers together and the picture is clear. The intent is universal. The maturity is not. And the problem usually concentrates in exactly one place: the step from the demo that shines in the meeting to something that runs every morning without anyone thinking about it. In every other engineering discipline this step has a name. It is called **Operations**. In AI projects it has been underrepresented so far.

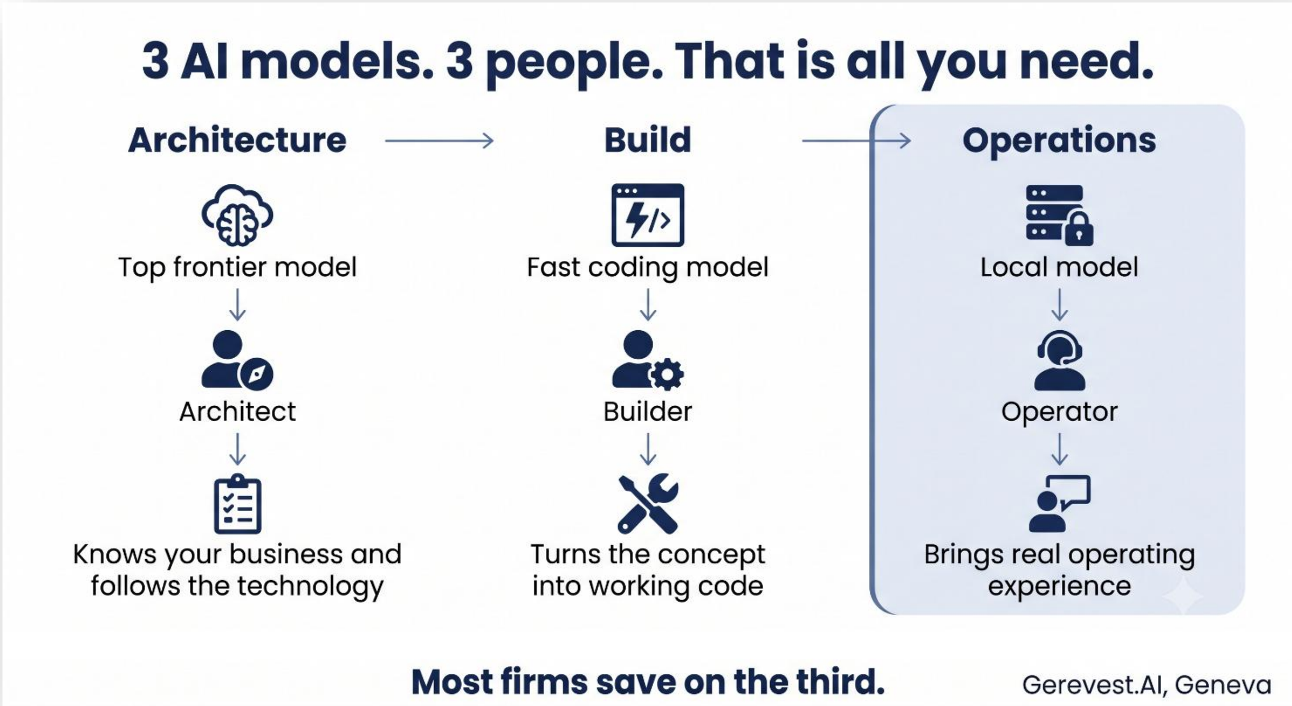


Figure 1. The Gerevest.AI Framework. Each phase of an AI project needs its own model class and its own competence profile. Source: Gerevest.AI.

2 An AI Framework for Efficiency and Autonomy

An AI project behaves like the construction of a new building. Someone designs it. Someone builds it. And then someone has to run it for years. Imagine you are planning a new watch factory in Geneva. Even the most reliable bricklayer will not be asked by the developers to first draw the blueprints for the new factory and then operate the production machines. The plans are made by an experienced architect who understands the requirements of a watch factory. And the new machines will still be operated by the trained watchmakers!

In AI projects at financial institutions, however, the following unfortunately happens all too often. A particularly talented working student, a new hire who still has to prove themselves, or someone from the existing IT without a real understanding of the business processes «builds» something from which management expects to save hundreds of thousands, if not millions of francs. Unfortunately, this method will not give you a future-proof watch factory. Nor an independent wealth manager.

You need an architect, builders and your own people.

It is much the same with the AI models as it is with the people. Why not one model for all three phases? Because the required properties are mutually exclusive. The architecture model must be the strongest on the market. That makes it expensive, and it is the one you can least host yourself. The operations model must be inexpensive, configurable and locally operable. That rules out the frontier. The build model must be specialized in coding and optimized for efficiency. The other two are usually far from it. Every attempt to press the three into one ends the same way: a beautiful proof of concept that is too expensive in operations. Or a production system that is already outdated in three months.

Architecture: The Frontier Model and the Architect

You use the strongest available model for a few hours. For thinking, not for producing. How does the process really run? Where does the data sit, in what form, and who owns it? What breaks when a format changes? What will the auditors ask? The token costs are worth it here. The leverage is the largest in the whole project. A bad architecture decision in week one is paid for every month after. Still, the frontier is very expensive. There are days when I pay 10 USD for 10 minutes with Claude Fable. In continuous operation that would be half a million a year.

So what should the competence profile of the human architect look like? The AI architect is usually not from «classical» IT. It should be someone who knows the business well enough ***and*** follows the technology closely enough to know what is realistically achievable. Large firms hire a Chief AI Officer. But for a small or mid-sized house or corporate team, that is usually too expensive. This is where targeted external support makes the most sense. Name someone from your business and equip this person with an external sparring partner.

Build: The Fast Coding Model and the Builder

Now volume counts. Building an agent means thousands of iterations. You optimize for speed and price per token, and you work via an API. In this phase, the strong open coding models from the Far East are now seriously competitive with the US models. When I first built software with AI coding in February 2025, I was 80% thrilled with the result. It looked good but it was not running reliably yet. A year later, February 2026, I was already 98% satisfied. The rest I adjusted manually. This is where the technology has advanced the most.

One important warning: you must give AI coding to someone who has programmed themselves before. Otherwise the person in charge will not notice the AI's mistakes. AI can make mistakes, there are misinterpretations of instructions or context overflows. Only an experienced professional notices when AI is faking something. Yet the opportunities to build your own software with AI and free yourself from costly processes and subscriptions are enormous, and in 2026 this trend is only beginning.

Operations: The Local Model and the Operator

This is the AI model that runs every morning for the next few years. The individual task itself is narrow and can be precisely specified. Think back to the watch manufacture. The leather strap can be punched by one machine. Another mills the case. Neither of them creates a tourbillon. Only the human specialist can. Nor do you send your half-finished watch to Asia for one inexpensive process step, you keep everything in-house and automate locally.

This is exactly how you must think about local AI. Welcome to the world of small, specialized AI models trained for **exactly one** task, running on your own local infrastructure. This world has grown a lot in 2026, and we will look at it more closely. What does that mean in practical terms? AI translates, schedules appointments and pulls the interest rate forecasts for your board presentation from five research reports. It extracts the ESG water consumption figures from 200 annual reports. In 2024 this was still difficult, today it runs.

Humans can use AI as an extreme lever. With the right machines you produce more, and better watches, because you concentrate only on the high value creation. Cover a much broader investment universe, make personalized proposals to clients, spend less time on «nonsense». Ask your people what they would like to have «automated away» (and not their own job) and what they would then do with their time. If the answer is something that either brings you more clients or delivers a better product to your clients, start the project.

So the Operator remains one of your own people. Someone who knows what a correct output looks like. Who notices when the quality is off. It is the person who owns the process today. The one who currently works and checks by hand, and therefore knows better than anyone what a wrong output looks like. What she lacks is a few weeks of structured training and someone who provides her with the newest machine. Tailored to exactly your process and no off-the-shelf AI. Operations includes control. The Operator checks samples, documents deviations and stops the process when the output drifts. Exactly these records are what Finma wants to see (Chapter 5).

Most firms save on step 3, Operations. Yet this is exactly where AI investments pay off in the long run. This is where your competitive advantage is created. This is where you secure your autonomy. And your jobs.

AI CHAMPIONS PROGRAMME — **from CHF 30'000 / 3 months**

We train two of your own people, build two real workflows together, and the playbook stays with you. Your operators, your infrastructure, your process. You pay Gerevest once, not forever.

Enquiries: info@gerevest.ai · Book an appointment online

3 Which Models, in September 2026

The latest model names change every few months. But the three model classes in our framework remain.

For **Architecture**, take the **top frontier**, via API: Claude Fable 5, Claude Opus 5, GPT-5.6 Sol, Gemini 3.1 Pro or Kimi K3. Kimi K3 is currently the only top frontier model that you could theoretically host yourself (see below); with all others you would send your client data through third-party data centers.

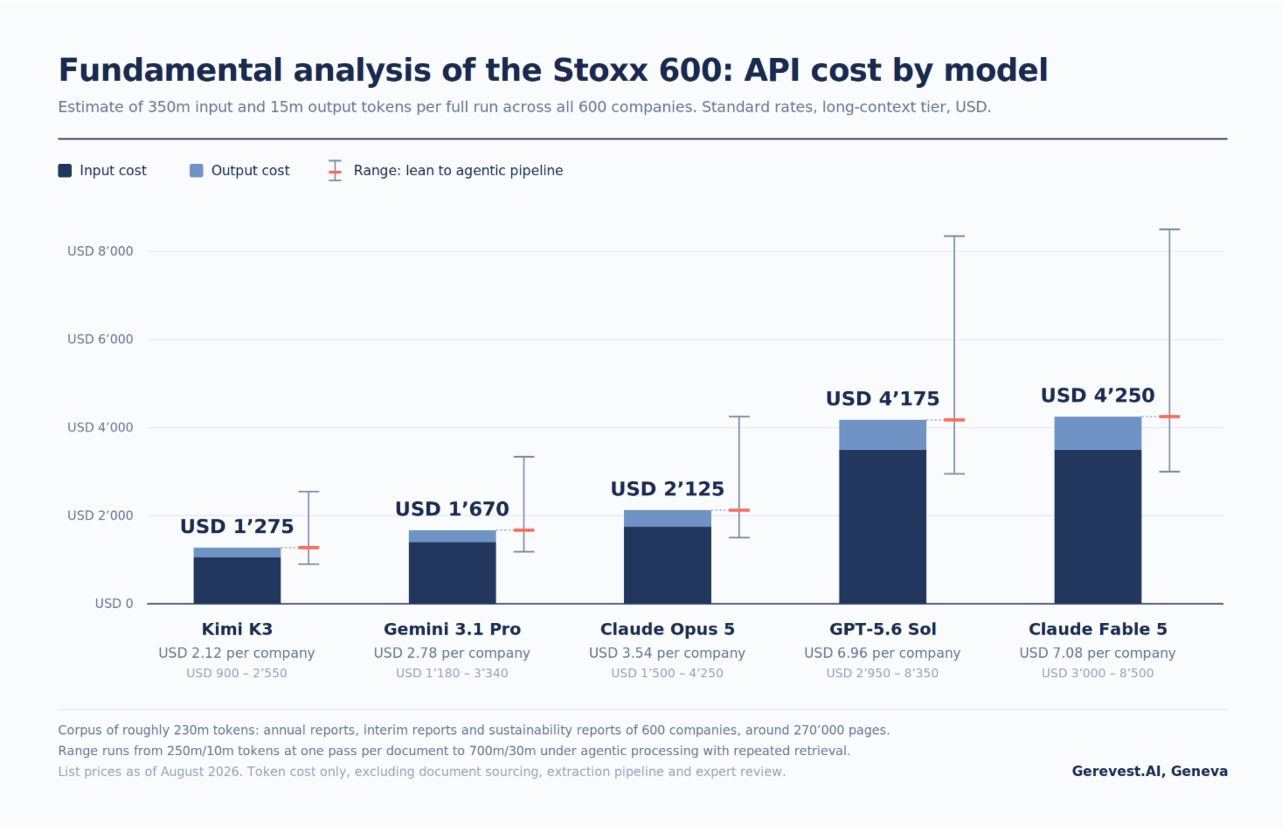
For the **Build** an efficient **coding model**: Also via API: Claude Sonnet 5, GPT-5.6 Terra, Gemini 3.6 Flash and of course from China: Kimi K2.7-code, MiniMax M3, GLM 5.3, DeepSeek V4 Pro, Qwen 3.8. Likewise, please do not use confidential client data while programming. In this league, however, there are already more and more models that you can run yourself on a modern laptop.

For **Operations** a small local model: Now it gets interesting, because in the broader market, there are many players specializing, e.g. Apertus from Switzerland and Mistral from Paris. Aleph Alpha from Germany and Cohere from Canada even merged recently. Since Google released the Gemma 4 «edge» series of AI models that run particularly well on Apple hardware, Mac Minis and Mac Studios have gone up in price and are often sold out for weeks.

Therefore, for locally operated processes on sensitive data: use small models.

But what about public financial data? Here you could also run the operations via the API.

Let us calculate a concrete price example: the fundamental equity analysis of all names in the pan-European Stoxx 600. Annual reports, quarterly reports, ESG – 600 names.



So for a four-digit USD amount in API costs, an AI analyst analyzes the entire European equity index for you. But if you want to ensure maximum quality, put one or two humans in front as Operators. It is like the watch factory. With two well-trained equity analysts + professional AI tools you can cover the whole market, like the big players. A huge opportunity for independent wealth managers.

To summarize, for sensitive data: a designated human remains as the Operator, supervising a locally operated, specialized small model. With public data you have the choice between a frontier model via API (with cost control) or, again, a small, specialized model. Similarly, a human remains in the loop as the Operator, whose expertise is augmented by AI. A single analyst can thereby cover considerably more names.

4 The Infrastructure Gap

The theory is clear, but what does it mean in practice? What hardware do I need to be truly independent? As a mid-sized bank, in August 2026 you would have to invest around half a million to keep up at the top. As an independent wealth manager, a few thousand francs are often enough. In between, there is a gap.

Why is that? Because research and development in 2026 concentrates on two extremes. Small, highly efficient models on consumer hardware on one side. Google's Gemma4 and Apple's Mac Mini / Studio are state of the art here. And thanks to the Alibaba-backed Chinese lab Moonshot AI, since July 2026 you can even run a true frontier model on your own hardware with Kimi K3. Theoretically.

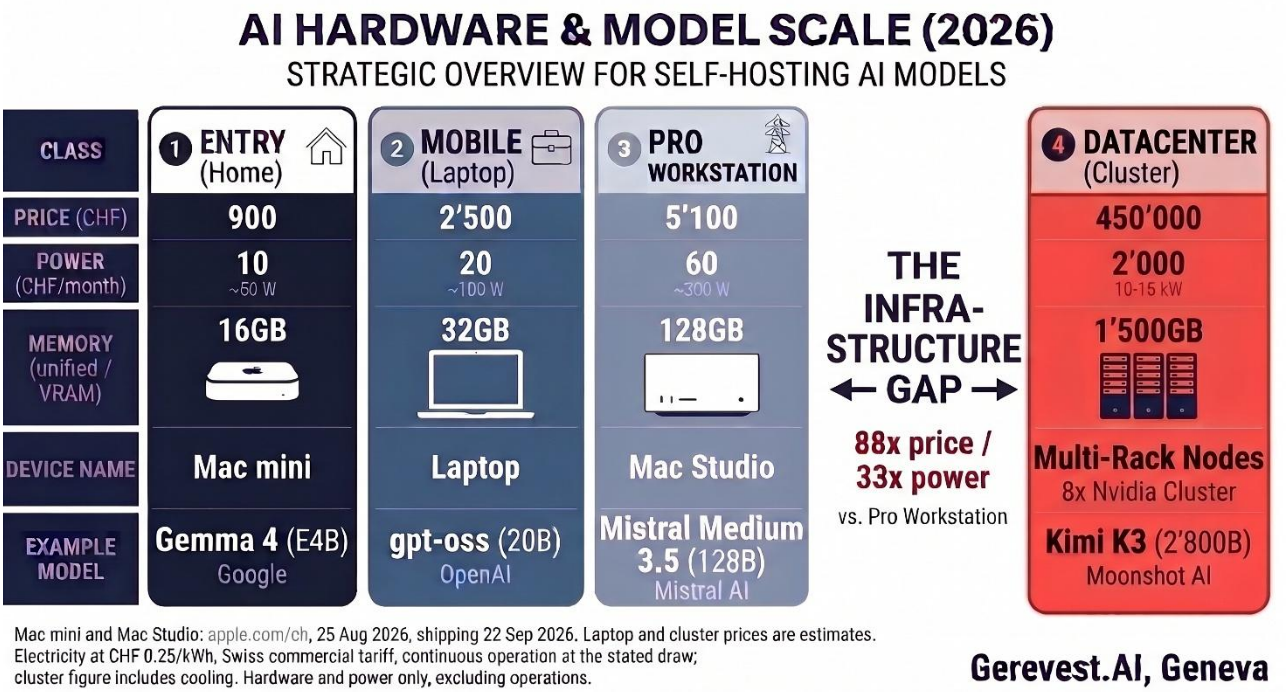


Figure 2. Purchase and operating costs for self-hosting. Price research and estimates as of August 25, 2026. Source: Gerevest.AI.

Since at my office location in downtown Geneva, running a local Kimi K3-capable data center would strain the available space, power supply and above all the cooling, I continue to use the model, in competition with Claude Fable 5, as pay-per-token via API. But locally on the small devices, the always-on, quiet, energy-saving Mac Mini, the modern laptop with sufficient memory, that is where the small models run smoothly.

The operating math is short: after the purchase you pay for electricity, around 60 francs a month for the Mac Studio in continuous operation (Figure 2). The ESG extraction from 200 annual reports then runs without a token bill. So two things speak for local hosting: data protection and price.

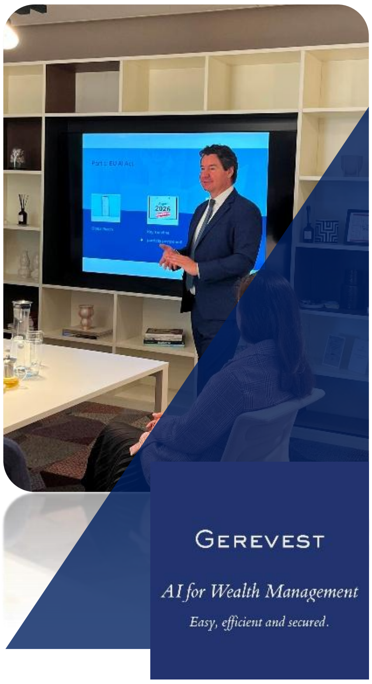
In July 2025, I built a Japanese equity market analyst with n8n and Mistral and waited up to four minutes on local infrastructure for the answer to a single prompt. That has been history since April 2026. The edge models answer locally in seconds.

For the use of small models, you can also rent capacity in Swiss or European data centers. For our own and for client purposes, we currently rent servers in Bavaria and in a suburb of Geneva, right next to the modern watch factories.

Conclusion: In 2026, the data security argument as an obstacle to AI adoption is obsolete. While running high-end models on your own infrastructure remains complex and mostly reserved for larger organizations, the answer for independent wealth managers and entrepreneurs is pragmatic: invest in powerful notebooks instead of tablets and smartphones. Combine this with targeted training for local AI applications. Remove the security risk of using private AI accounts for client communication. Your client data stays on your employees' laptops.

5 Regulation: What Really Changed on August 2, 2026

We have now covered theory and practice. What is missing is regulation.



Since August 2026, Article 50 of the EU AI Act applies, the transparency obligation. It also applies to Swiss wealth managers and entrepreneurs as soon as your systems or content reach clients in Germany, France or other EU countries. Systems that interact with humans must disclose that they are AI.

If you publish AI-generated text, it must be labeled as such, unless a human has reviewed it and carries editorial responsibility. The penalties for violations of the AI Act can be very high: up to 15 million euros or 3% of worldwide annual turnover. Therefore, include AI labeling in your next disclaimer if you generate content, such as infographics, using AI.

Postponed to December 2, 2027, are the Annex III obligations for high-risk systems. Examples include automatic credit scoring or AI systems in the HR department.

The obligation to train your employees, however, has been in force since February 2025. «AI literacy» is the keyword, and many companies already train regularly. We recommend seeing this as an opportunity, not a tiresome duty. Employees trained in AI competence take your company further.

Standard training programs often fall short of the standards required by regulatory bodies and internal audits. A training is only effective if it is tailored to the specifics of your industry and your particular needs. What matters is that the enablement starts where AI is actually applied in day-to-day business. Returning to our watch factory analogy: for the vision of the Architects and Builders to work in practice, the Operators need the corresponding AI literacy.

In summary: The AI Act is in force and, with EU clients, also applies to you as a Swiss company. The penalties are high. But no panic. Label AI when your clients interact with it or receive AI-generated content. And train your employees regularly, above all the Operators.

And what does Finma require?

While the EU is phasing in its requirements, the binding Swiss regulations have been in effect since December 2024. Finma Guidance 08/2024 describes what supervised institutions should demonstrate when using AI. It is technology-neutral and principles-based. At its core it requires five key elements, and all five have a place in our framework.

FINMA REQUIRES	IN THE FRAMEWORK THIS IS
A central inventory of AI applications, with risk classification and the measures that follow from it	Operations
Named responsibilities across development, implementation, use and monitoring	Architect, Builder and Operator (+ Compliance)
Defined model tests and the system controls behind them	Operations
Documentation standards for all of it	Operations
Training of those involved, including where the work is outsourced	The literacy obligation, from the other direction

AI LITERACY TRAINING — **from CHF 5'000 / day**

One day, up to ten participants, role-specific for investment, risk, operations or sales. Including handout and documentation of the learning content. Adapted to your company and the relevant topics of the financial industry.

Enquiries: info@gerevest.ai · Book an appointment online

6 Your Next AI Workflow, Ready by Christmas

OCT '26: ARCHITECTURE	NOV '26: BUILD	DEC '26: OPERATIONS
• Engage the Architect	• Engage the Builder	• Name the Operator
• Choose a workflow	• Build the workflow with a fast coding model	• Run the workflow locally or via API
• Measure the time (legacy)	• Measure the time (new)	• Test run & go-live

In three months, you get a workflow with real experience and a measurable result for your board. That is considerably better than three pilot projects – or a long list of planned pilots.

SIX QUESTIONS BEFORE YOUR NEXT AI PROJECT

1 Can you name the person who will run your AI workflows in twelve months?
2 Do you know which of your systems interact with clients, and do they disclose that they are AI?
3 Is your AI literacy obligation documented, or merely done?
4 Do you know what happens to your AI bill if usage triples?
5 Could you switch model providers within a quarter without rebuilding?
6 Does your data reach the model in a structured format, or is someone still copying it by hand?

Fewer than four clear yeses, and you have a pilot, not operations.

ABOUT THE AUTHOR

Dr. Andreas K. Janoschek is the founder of Gerevest.AI, a Geneva-based AI boutique that supports banks, insurers and wealth managers in the secure and sovereign implementation of artificial intelligence. After completing his doctorate in mathematics on learning algorithms, he worked as a portfolio manager at Allianz Global Investors, Munich Re, and as Chief Investment Officer of a life insurer.

You will meet him along the axis between Zurich and Geneva, wherever AI meets the financial industry.

This framework grew out of 24 months of AI project experience, from our own projects as well as client projects, and from ongoing exchange with independent wealth managers, private banks, asset managers and entrepreneurs.

Where Would Your Three People Sit?

A paid hour filters better than a free meeting. Book an Executive AI Coaching session. You will leave the conversation with a clearly defined profile for your Architect and your Operator – and a conceptual workflow that justifies the investment. Literacy Training and Champion Programme follow afterwards, if it fits.

Gerevest.AI · info@gerevest.ai · In person in Geneva and Zurich or hybrid across Europe · Book an appointment

Insights

SOURCES

[1] The adoption figures come from the study «AI Use Cases in External Asset Management (EAM)» by the Lucerne University of Applied Sciences and Arts (IFZ) with Finup, commissioned by Innosuisse, published on August 3, 2026: 41 firms surveyed, nine interviews, against the background of 1'427 licensed Swiss wealth managers; reported among others on investrends.ch and finews.ch. The regulatory information is based on the EU AI Act as amended by the Digital Omnibus Regulation (EU) 2026/1744 (in force since July 27, 2026) and the FINMA Guidance 08/2024 of December 18, 2024; the model and hardware figures are based on manufacturer documentation and public benchmarks as of August 25, 2026, API prices are indicative US list prices excluding taxes. The API cost estimates are based on our own experience and projections made with the help of Claude (Anthropic); they are for illustration only and may be higher or lower in your individual use case. The hardware prices are based on internet research at Swiss online retailers and may change. This document informs; it is neither legal advice nor an investment recommendation. Any liability is excluded.

This document was created by humans with the support of artificial intelligence. For the infographics, formatting, linguistic corrections and translations we used, among others, tools from Anthropic (Claude), Google and open-source tools. Editorial responsibility and the substantive design lie entirely with the author and are based on our own assessments and experience.

The basic idea of this framework was presented by Andreas Janoschek on July 28, 2026 in his column on finews.ch; the present elaboration was written in August 2026. The Gerevest prices quoted in this document are introductory prices, valid until December 31, 2026, and exclude VAT.

Photos: © Andreas Janoschek (AI for Good, Genève; Rue de la Cité 1, Genève; Public Talk «AI in Action», Gland; Public Workshop, Zürich).