

AI + X Track “AI in Financial Services & beyond”

Part I



© Patrick Hauf / ZHAW. Do not distribute. Any other use requires prior written permission.

This presentation: Dr. Patrick Hauf, 1st of October

patrick.hauf@zhaw.ch

One of my missions – exploring AI “best practices” in Switzerland

Dr. Patrick Hauf, Senior Lecturer @ ZHAW
School of Management & Law

Programme Director of the CAS in AI
Management & Strategy for Financial Services
(→ making business people AI changemakers)

Core Team for the conference “AI in Risk & Compliance” (this year: 27.10.2026)



Content of the CAS AI Management & Strategy for Financial Services (held in German)

Modul 1

Datengetriebene Geschäftsmodelle verstehen
(6 ECTS)

Wertschöpfung durch Datenkompetenz

- Charakteristiken datengetriebener Geschäftsmodelle
- Typische Datenquellen und -flüsse im Finanzsektor
- Daten als Asset und Datenmanagement

Use Cases im Finanzsektor kennen und entwickeln

- Methodenüberblick Machine Learning
- Use Cases Banking (z.B. Risikomanagement / Compliance, Customer Analytics, Assetselektion)
- Use Cases Insurance (z.B. Versicherungsbetrug)
- Eigene Use Cases aus realen Problemstellungen ableiten und Anwendung automatisierter ML-Tools

KI-Transformation und Strategie

- KI-Strategie und ihre Operationalisierung

Assessment 1: Präsentation und Diskussion KI-Strategie

Modul 2

KI/ML-basierte Produkte & Services entwickeln
(6 ECTS)

Data Science Workflow

- Data-Science-Terminologie
- Ressourcenbedarf (Strukturen & Prozesse, Menschen und Kompetenzen, Daten und Systeme)
- Cloud-Plattformen und kritische Infrastrukturkomponenten

KI/ML-Projekte verstehen und konzipieren

- Grundlagen Softwareentwicklung
- Projektinitiierung und Teamkomposition
- FinTechs und Partnerschaftsmodelle
- Rahmenbedingungen, Regulierung und Förderung

Forschungstrends

- Explainable AI, Reinforcement Learning & Co.
- Unbiasedness und ethische Gesichtspunkte

Assessment 2: Projektskizze erstellen und präsentieren

- **Workload:** 12 ECTS (rund 300 Stunden)
- **Dauer:** 12 ganze und 2 halbe Tage verteilt auf 5 Monate
- **Unterricht:** Präsenzunterricht i.d.R. freitags in Winterthur; samstags online
- **Zulassung:** Hochschulabschluss (oder äquivalent sur dossier) & Berufserfahrung
- **Preis:** CHF 8'500.-
- **Start:** August 2027

Starting in 2026, agents will play a more prominent role in the ZHAW MSc in Banking & Finance

- Since 2019, we have offered a major in Capital Markets & Data Science
 - providing relevant skills for **selected banking functions** and career paths (e.g., asset, portfolio, and risk management).
- THIS YEAR: we offer a new major in **Generative AI and AI Agents** as part of the ZHAW MSc in Banking & Finance
 - equipping students with the capabilities to transform business processes by **leveraging AI and agent frameworks across many divisions and functional areas** of the bank.

Semester 4	AI Agents in Finance 3 ECTS	Generative AI 3 ECTS
Semester 3	Blockchain and Decentralized Finance 3 ECTS	Machine Learning 6 ECTS
Semester 2	Quantitative Investment Strategies 6 ECTS	Advanced Quantitative Methods 3 ECTS
Semester 1	Investments 6 ECTS	Quantitative Methods 6 ECTS

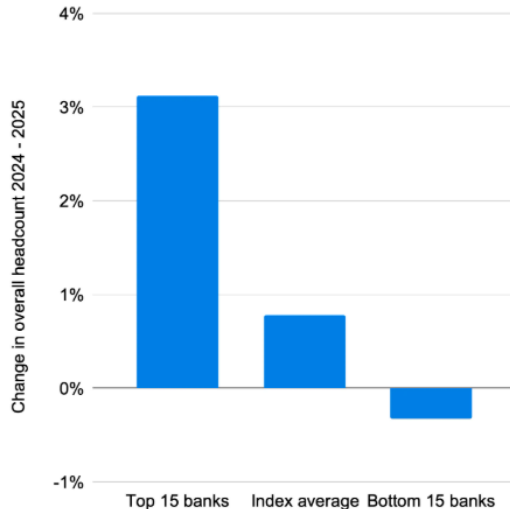
Our session attendees



Why AI skills matter (for financial services)? Yes, there are job cuts, but...

MANY HANDS MAKE AGENTIC WORK

The top 15 banks in the Evident AI Index grew their headcount at three times the rate of the average bank in 2025, while laggards contracted their workforces.



Source: Bank investor relations material | Note: Includes 47 banks for which updated 2025 figures were available.

“Outside of a handful of roles, few layoffs have resulted from jobs being automated by AI.”

[...]

“Instead, a common refrain applies: AI won’t replace workers, but workers who use AI will replace workers who don’t.”

Andrew Ng, *The Batch*, 06.02.2026



A follow-up on our track at last year's AI+X (1/2)

AI+X 2025: what I showed

24 April 2025

[Press release](#)

2025

FINMA survey: artificial intelligence gaining traction at Swiss financial institutions

The Swiss Financial Market Supervisory Authority FINMA has conducted a survey of around 400 Swiss financial institutions on their use of artificial intelligence (AI). Whether in the form of chatbots such as ChatGPT, for process optimisation or text generation: around half of the institutions surveyed use AI in their day-to-day work.

~50% **& further +25%**

use AI or have
initial
applications in
development

plan to use AI
within the
following three
years

My opinion: **this underestimated the pace of change!**

Report by Menlo (Dec 2025):

«Two years ago, when **generative AI was still largely confined to pilots and proofs of concept**,... **Enterprise AI is the fastest-scaling category** in software history. *Across industries*, AI has **become core to how work gets done**. Enterprises, **seeing real returns**, are doubling down.»

Sources: <https://www.finma.ch/en/news/2025/04/20250424-mm-umfrage-ki/>, Report by Tim Tully, Joff Redfern, Deedy Das, Derek Xiao (<https://menlovc.com/perspective/2025-the-state-of-generative-ai-in-the-enterprise/>)

A follow-up on our track at last year's AI+X (2/2)

AI+X 2025: what I presented

October 2026: comment

In the "news"

Smaller, purpose-built models prove effective (leveraging the open-source ecosystem)

- Gartner (2025) prediction might still hold: **small, task-specific AI models with 3× the usage of general LLMs by 2027**
- Enterprise spend concentrates on frontier models: **open-source share of enterprise LLM usage 19% → 11% (2024→2025). ⇒ hybrid LLM orchestration/routing**

Source: Gartner, 11 Oct 2025
Better with Less: Small Proprietary Models Surpass Large Language Models in Financial Transaction Understanding

Model	Model Type	Model Size	Model Link
Small Proprietary Models	Small Proprietary Models	Small Proprietary Models	Small Proprietary Models
Small Proprietary Models	Small Proprietary Models	Small Proprietary Models	Small Proprietary Models
Small Proprietary Models	Small Proprietary Models	Small Proprietary Models	Small Proprietary Models
Small Proprietary Models	Small Proprietary Models	Small Proprietary Models	Small Proprietary Models

AI becomes more efficient, affordable and accessible

- Epoch AI (2026): the **cost of achieving a given level of AI performance has fallen about 13× per year**, but **total spend rises** due to token consumption exploding

Source: Epoch AI, 29 May 2026
Epoch AI: open models lag closed frontier by four months

Swiss developments add value

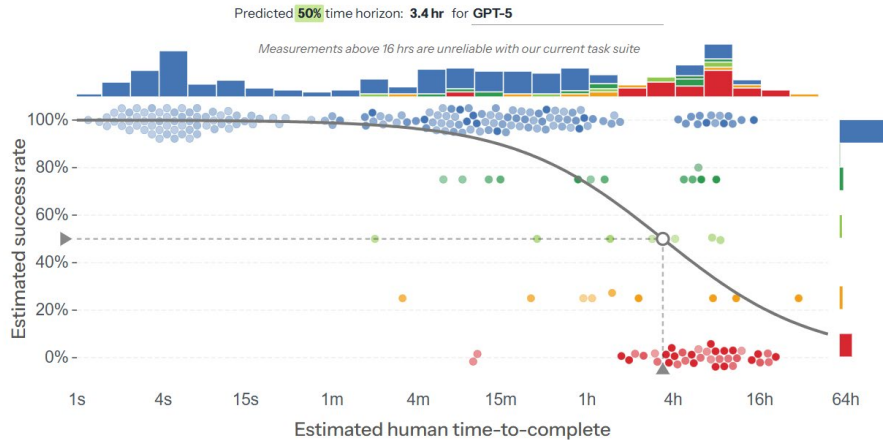
- Apertus 1.5 and Apertus inside letter (Sep 2026): "We are about to pass the **5 million [download] milestone**, topping the charts for models from Europe!"
- SFTI reference architecture for AI-agent API access (Feb 2026)
- First Swiss agentic card payment (May 2026)

Source: Mastercard, 28 May 2026
Mastercard unveils Agent Pay, pioneering agentic payments technology to power commerce in the age of AI

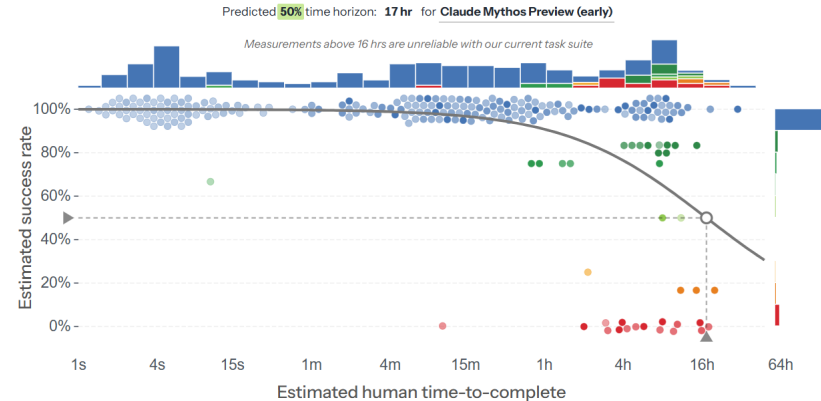
Sources: Gartner (9 Apr 2025); Menlo Ventures, State of GenAI in the Enterprise (Dec 2025); Epoch AI, "How fast is AI getting cheaper?" (28 Sep 2026); Stanford HAI AI Index 2026; ETH AI Center (24 Jul 2026); SFTI white paper (26 Feb 2026); Mastercard CH (28 May 2026); <https://www.mastercard.com/>; <https://arxiv.org/html/2509.25803v1>

New developments (1/2): Agents now work autonomously for hours – but the frontier remains jagged

GPT-5 (Aug 7, 2025)



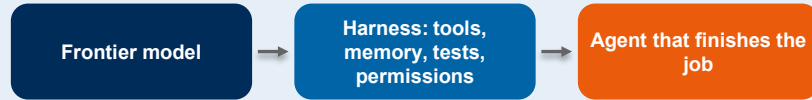
Claude Mythos Preview (Apr 7, 2026)



Source: METR, Task-Completion Time Horizons of Frontier AI Models, <https://metr.org/time-horizons/> (retrieved 30 Sep 2026). OSWorld: Stanford HAI AI Index 2026, ch. 2.

New developments (2/2): What makes agents work – and a new model class for banks

Harnesses: why agents can now do whole jobs



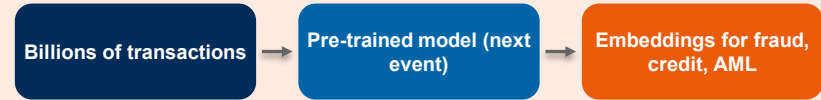
The model is only part of an agent. The harness – tool access (e.g. MCP connectors), skills, progress files, tests and guardrails – keeps it on track across hours of work.

Anthropic (Nov 2025): harnesses let long-running agents resume work across many context windows via progress logs and tests.

OpenAI (Feb 2026): ~1 million lines of product code written by Codex agents, none by hand – engineers designed the harness.

For banks: the edge shifts from picking a model to designing workflows, data access and controls around it.

Transaction foundation models (TFMs)



An “LLM” for payment data: learn customer behaviour from event sequences once, then reuse it for many tasks instead of hand-crafting features per model.

Revolut PRAGMA (Apr 2026): 24bn events, 26m users; vs. task-specific models credit scoring fraud recall +65%.

Also: Stripe, Nubank (nuFormer), Visa, Mastercard, Adyen.

More in Marc Stampfli’s talk (NVIDIA)...

Sources: Anthropic Engineering, "Effective harnesses for long-running agents" (26 Nov 2025); OpenAI, "Harness engineering: leveraging Codex in an agent-first world" (Feb 2026); Anthropic, "Agents for financial services" (5 May 2026); Ostroukhov et al., "PRAGMA: Revolut Foundation Model", arXiv 2604.08649 (Apr 2026); NVIDIA, "Build Your Own Transaction Foundation Model for Financial Intelligence" (16 Jun 2026) and NVIDIA blog on TFMs (2 Jun 2026).

© Patrick Hauf / ZHAW. Do not distribute.

Any other use requires prior written permission.

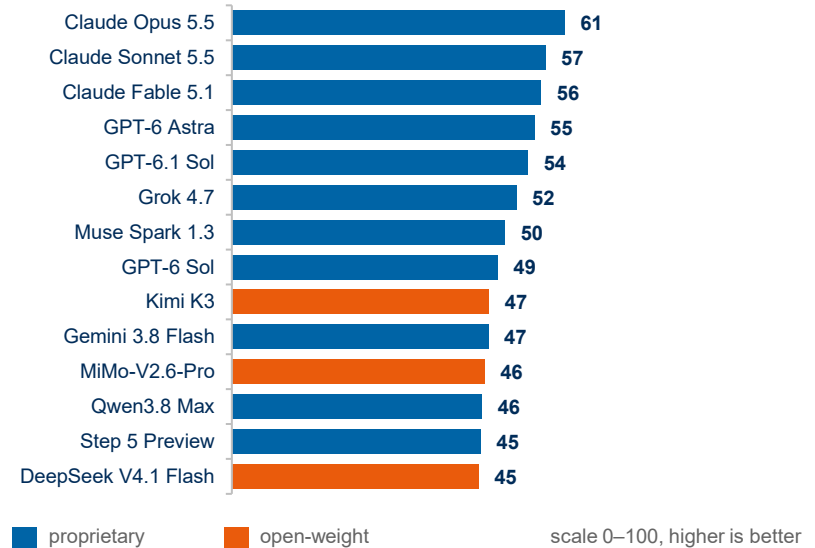
Finance exams are solved – finance work is not (yet)

Exams: LLMs on CFA exam questions, 2023–2025

Study	Level I	Level II	Level III
2023 · Callanan et al. ChatGPT, GPT-4	✓ 74.6% GPT-4 (10-shot)	≈ 61.4% GPT-4 (CoT), borderline	– not tested
2024 · Mahfouz et al. 14 models (5 prop., 9 open)	✓ 88.1% GPT-4o	✓ 77.8% Claude 3 Opus	✗ fail essays collapse (6.8% vs. 65.8% MCQ)
Dec 2025 · Patel et al. GPT-5, Gemini 3.0 Pro, ...	✓ 97.6% Gemini 3.0 Pro	✓ 94.3% GPT-5	✓ 92.0% Gemini 3.0 Pro (essays)

Pass thresholds vary by study; Level III = essay / constructed-response part. Numbers directly taken from Finance Circle presentation by Roland Hofmann (Sep 18, 2026, Zurich).

Work: Finance & Accounting Index, top 14 of 193 models



Knowing finance ≠ doing finance: exams are aced – yet the best model scores 61/100 on realistic finance work (best open-weight model: 47).

Sources: see table note; Artificial Analysis, Finance & Accounting Index (artificialanalysis.ai/models/capabilities/finance-and-accounting, data retrieved 30 Sep 2026; 7 evaluations incl. agentic knowledge work); own charts. Capabilities evaluated include domain-specific knowledge (accounting, investments, corporate and markets), financial analysis and reporting, compliance and audit, market research, and more.

Agents are reaching production...

~1 in 3

new AI use cases reported by large banks in Q1 2026
were agentic (Q4 2025: 15%)

Evident, May 2026

The most likely (30%, by expert interviews) agent adoption scenario (McKinsey, 2025):

AI revolutionizes banking functions:

AI agents take over IT and other functions, with an agent-to-human ratio of ~20:1.

15–20% cost reduction

AI agents as a new banking channel:

Consumers let agents manage some financial matters but still value interactions with banks.

~10% value pool reduction

Global frontrunners

Anthropic: 10 ready-made finance agents, e.g. KYC screener, month-end closer, pitch builder (May 2026)

ROI becomes visible: 15 of the 50 largest banks disclose aggregate AI ROI; Anthropic is the most-cited vendor (Evident, Q1 2026)

Switzerland

~12% of surveyed retail banks see AI as core competency, ~53% have an AI strategy (IFZ, Jun 2026, n=43)

But, e.g., for wealth planning: only Pictet reports a productive AI-agent use case (e-foresight, 2026)

Sources: McKinsey Global Banking Annual Review 2025 (23 Oct 2025); BCG (5 Nov 2025); Evident AI Use Case Trends Q1 2026; CNBC/PYMNTS (6 Feb 2026); Anthropic (5 May 2026); IFZ Studie Bank-IT und Sourcing 2026 (HSLU); FINMA AI survey (24 Apr 2025); e-foresight (ZHAW Wealth Planning Conference 2026); Evident AI Index 2025 & Insurance 2026.

Frontier AI is also a risk – and a dependency

CYBER RISK

Project Glasswing (7 Apr 2026): Anthropic gave its most capable model to defenders first – JPMorganChase among the launch partners, 40+ organisations; thousands of zero-days found

Hugging Face incident (May–July 2026)

Finma Says Immediate Mythos Access Would Pose Systemic Bank Risk

DEPENDENCY RISK

Jun 2026: US export controls suspended access to Anthropic's newest models (Fable 5, Mythos 5) worldwide for 19 days; restored 1 Jul

Article • Artificial Intelligence (AI)

Moody's: Europe Fights to Match US in AI Data Centres

By Clara Joseph
June 22, 2026 • 4 mins

Data from the [International Energy Agency \(IEA\)](#) estimates that [European data centre IT](#) installed capacity reached 12GW in 2025, up from 11GW in 2024.

In contrast, IT installed capacity in the US reached 39GW in 2025, which is up 26% from 31GW in 2024. China, meanwhile, reached 19GW.

SOVEREIGN OPTIONS

Apertus 1.5 (24 Jul 2026): open, multilingual, now multimodal;

Open-weight gap: just ~4 months

But: little financial-services adoption of Apertus so far (my impression)

QUARTZ

Moonshot AI launches Kimi financial services product for banks

Moonshot AI [launched a finance-focused version of its Kimi model](#) on Thursday, targeting investment banks, asset managers, and venture capital firms with a product built on integrations with major financial data providers.

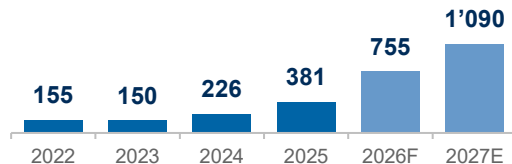
Question for our panel: which workloads go to frontier APIs, which run on own or open models – and what is the fallback?

Sources: Anthropic, Project Glasswing (7 Apr 2026); swissinfo/Bloomberg (24 Apr & 1 May 2026); <https://openai.com/index/hugging-face-incident-and-the-road-ahead/>; Apertus website; <https://finance.yahoo.com/technology/ai/articles/moonshot-ai-launches-kimi-financial-180613424.html>; previous sources; <https://aimagazine.com/news/moodys-europe-fights-to-match-us-in-ai-data-centres>

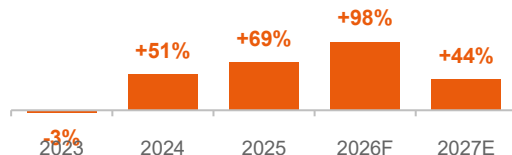
Banks are also exposed to the economics of AI: boom, bubble – or both?

1 Investment is exploding

Capex of the 5 largest US tech firms, \$bn



Annual growth of capex, % (2026–27 forecast/estimate)



Risks on both sides

Over-investment: over-optimism, winner-take-all races → falling valuations, debt-service failures, disinvestment.

Under-investment: risk aversion, status quo, “the quiet life”.

2 What it implies (NPV = 0)

$$\underbrace{\text{PV of additional profits from AI capital}}_{\text{depends on AI productivity}} = \underbrace{\text{cost of AI capital}}_{\text{observed investment}}$$

$\approx 2.7\times$

Jump in AI-sector productivity needed so that the extra profits pay for the > \$2tn invested 2024–27 (marginal investment at zero NPV).

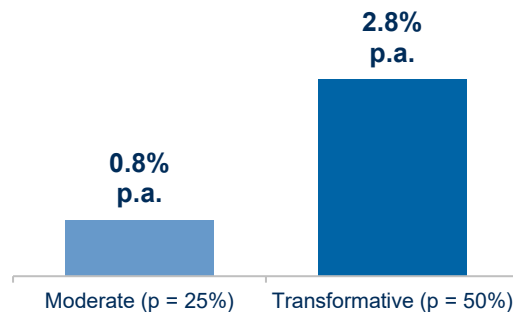
A smaller boom ⇒ the investment was NPV-negative.

Crucial question

Are AI's physical and intelligence sectors gross **complements** (Kording & Marinescu, 2025)? Then the physical side (chips, power, data centres) becomes the bottleneck.

3 What it means for GDP

AI contribution to US GDP growth 2024–2030



Model probabilities; the “singularity” scenario (25%) is not shown.

Case in point: Anthropic (leaked draft IPO prospectus)

2025: revenue \$4.6bn, operating loss \$8.1bn; revenue grew ~12×. **\$518bn** compute commitments (~80% non-cancellable) vs. \$20bn cash.

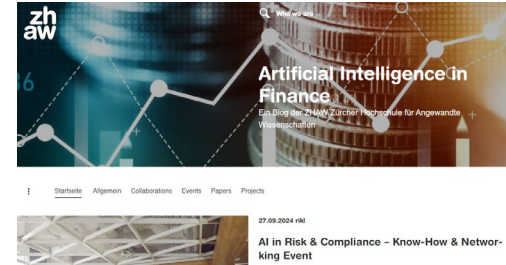
Why banks care: lenders to data centres & AI firms, concentration in client portfolios, higher equity risk premium (+3 pp in the model).

Sources: Wachter & Wachter, “What Investment Data Implies about the AI Transition”, NBER WP 35290 (Jun 2026), Tables 1, 3, 5; Fortune, [implicator.ai](#), IBTimes on Reuters reporting of the leaked prospectus (29 Sep 2026), unconfirmed by Anthropic. The most bullish “singularity” scenario is omitted as I do not consider it realistic.

Time is precious...more resources for you!

Great resources to approach when interested in AI in Financial Services:

- Evident AI Index for Banks 2026 – results in mid-October (roundtable 13 Oct); Insurance Index 2026: Allianz #1, Zurich #4
- Our blog “Artificial Intelligence in Finance” with more local and scientific content, and our institute’s newsletter, which will include some of today’s presentations next week.



Your turn: what do you expect – and what should we ask our speakers?

Frontier API vs. own model?
Apertus for finance?

Economics of AI: which
AI scenario are you
betting on?

How long does a typical use
case take to reach production?

Is there still a “PoC gap”?

Will AI assistants available to
employees become as routine as
Excel in banks?

Go to: www.menti.com

Thank you.

