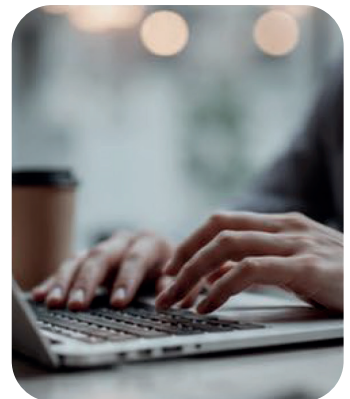
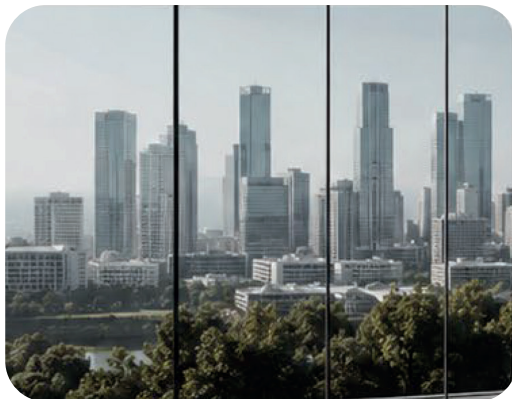


# *The new European frontier of B2B AI*



## Strategic Intelligence Report

Seizing the opportunity

# The new European frontier of B2B AI

---

Strategic Intelligence Report

# The new European frontier of B2B AI

## Executive Summary \_\_\_\_\_ 4

### 01 **Scientific, innovation and talent ecosystem** \_\_\_\_\_ 8 *From research rigour to internal organisational capability*

- **Scientific ecosystem:** The global footprint of financial AI, the speed of innovation and the map of institutional competitiveness \_\_\_\_\_ 9
- **Innovation ecosystem:** The European supernodes of excellence (Paris, DACH, London, Spain) and Lab-to-Market transfer \_\_\_\_\_ 20
- **Transfer, talent and governance:** New hybrid profiles (Analytics Translator), operational restructuring in Fusion Teams and strategic retention \_\_\_\_\_ 27

### 02 **Technological evolution, models and infrastructure** \_\_\_\_\_ 36 *From consumer AI to operational and sovereign efficiency*

- **From mass AI to industrial AI:** the transition towards industrial AI (SLMs), the Build vs Buy financial analysis, and the control of Total Cost of Ownership (TCO) \_\_\_\_\_ 37
- **Control architectures and the agentic frontier:** The Sovereign RAG paradigm, Zero Egress networks, the leap to system 2 and automation through agents \_\_\_\_\_ 45
- **Critical infrastructures and sustainability:** The European Gigafactories, the Spanish supercomputing hub (RES) and the imperative of Green AI \_\_\_\_\_ 53

### 03 **Regulatory framework and compliance** \_\_\_\_\_ 62 *The regulatory framework as a strategic defensive moat*

- **Regulation as a product:** Compliance by Design and the competitive advantage over non-auditable models \_\_\_\_\_ 64
- **Integrated management and MRM:** AI within the Model Risk Management (MRM) framework for controlled scaling \_\_\_\_\_ 65
- **Current case:** access to frontier models and operational sovereignty \_\_\_\_\_ 66

## Strategic Intelligence Report

---

# The new European frontier of B2B AI

---

## Executive summary



The report analyses the stage of maturity that foundation AI is undergoing in Europe and its potential impact on critical sectors. The first wave of adoption has been marked by large general-purpose models and consumer applications, but the strategic debate is beginning to focus on more specialised, secure and governable uses. This change is particularly relevant for critical and regulated sectors, where any technological deployment must coexist with strict requirements for traceability, resilience, privacy and risk control.

For the banking sector, AI represents an opportunity for transformation, although its value will depend on the ability to integrate it in an orderly way into real business processes. Priority should be placed on selecting the areas where it can generate efficiency, improve decision-making, strengthen operational resilience, and protect data sovereignty. That prioritisation requires combining technological ambition with regulatory discipline, economic judgement and a governance architecture prepared to scale.



# From the generalist model to sovereign AI in critical sectors

*The advantage lies not in the size of the model, but in its secure integration into critical processes.*

## Industrial, regulated, and efficient AI

The analysis confirms that the United States and China maintain a structural advantage in private investment, general-purpose foundation models, and cloud infrastructure. Faced with this asymmetry, Europe is shaping a different approach, one more focused on vertical specialisation, regulatory trust and control of industrial data.

This position is particularly well suited to banking. In an activity where trust, data protection and risk management form part of the core of the business, the adoption of AI cannot rely on opaque models or on architectures that leave sensitive information outside the control perimeter.

The priority must be AI that is understandable, auditable and governed by design, capable of operating in secure environments and aligned with European regulatory requirements.

The report focuses on **three key ideas**.

The first is that **specialised models** can offer a much more favourable balance between cost, latency and operational control. The second is that **sovereign RAG\*** makes it possible to separate the AI engine from the corporate memory, preventing internal knowledge from being absorbed by external providers. The third is that **regulatory compliance**, far from being merely a burden, can become an industrial quality benchmark and a competitive barrier against less transparent or difficult-to-audit technologies.



*The US mobilised \$109 billion in private AI investment in 2024, compared with \$19.42 billion in Europe. The European response cannot be based on replicating that scale, but rather on combining selective investment, sovereignty, trust and B2B specialisation.*

\* Sovereign RAG for governed corporate knowledge.

# Main trends

Taken as a whole, the report points to five major trends. They are not isolated lines, but rather pieces of the same transition towards more efficient, more controlled B2B AI that is aligned with European regulatory requirements.

## **1. The era of “One Model Fits All” is beginning to run out**

Large generalist models will continue to be relevant, but they are not the optimal answer for all regulated business processes. For recurring, document-based or highly regulated tasks, Small Language Models (SLMs) and specialised models offer a more efficient alternative. They make it possible to reduce inference costs, improve latency and strengthen the competitive advantage through greater control over data, weights and traceability.

## **2. Sovereign RAG becomes a key architecture**

The separation between the model and corporate memory will be a structural decision. A regulated organisation does not need to retrain models using all its internal knowledge, but rather to connect controlled models with governed document repositories. This architecture makes it possible to use generative AI without losing control over case files, internal policies, regulatory documentation or customer data.

## **3. Regulation shifts from being a brake to a competitive advantage**

EU AI Act, DORA, GDPR and NIS2 raise operational requirements, while also

professionalizing the market. Entities capable of demonstrating traceability, human oversight, risk control, data quality and technological resilience will be better placed to scale AI in critical processes. In banking, rigorous compliance can act as a mark of trust.

## **4. Talent is no longer a support function; it becomes core infrastructure**

Competitive advantage will no longer depend solely on hiring specialised technical profiles. The report points to a hybridisation between business experts, legal profiles, risk teams, technologists and European scientific networks. AI-driven differentiation in banking will be built on domain and business knowledge, not solely on model engineering.

## **5. Maturity and control must guide prioritisation**

The report identifies RegTech, fraud prevention, AML, KYB, document AI, synthetic data, model governance and supervised agents as priority processes. These are areas where the Bank can capture efficiency, reduce operational risk, and build scalable capabilities without assuming excessive dependence on proprietary vendors.

*The key question is no longer which model to use. The relevant question is which processes must remain under sovereign control, which can be supported by external providers, and what level of oversight each automated decision requires.*

# Executive roadmap

The strategy emerging from the analysis of the report does not envisage a massive, undifferentiated transformation. It proposes a gradual adoption, with different priorities depending on the time horizon, level of technological maturity, and regulatory sensitivity in each use case.

## 1. Short term, from 0 to 12 months

### **Sovereign RAG, document automation and RegTech**

The priority in implementation should focus on cases of immediate operational impact.

Internal document assistants, governed semantic search, KYC/KYB automation, reduction of false positives in AML, and compliance copilots offer a favourable combination of maturity, return, and control. At this stage, it is advisable to implement the Sovereign Wrapper as a cross-cutting layer for access, auditing, logs, and data output control.

## 2. Medium term, from 1 to 3 years

### **Supervised agents and traceable reasoning**

The second phase should move towards systems capable of helping with more complex processes, not only in text generation. These supervised agents could support tasks such as credit analysis, alert management, case file review, treasury, and international trade finance or the prevention of fraud. Their use should always be governed by clear business rules, with human validation (maker-checker) before any sensitive decision and with a complete record of the steps followed by the system. In parallel, Quantum AI, optimisation,

and post-quantum cryptography (PQC) must remain a priority technical agenda item, with active monitoring and early PQC preparedness.

## 3. Long term, more than 3 years

### **Hybrid AI, financial simulation and post-quantum technologies**

The third phase must remain under active monitoring and controlled experimentation. It includes neuro-symbolic AI, digital liquidity twins, market simulation, Quantum AI applied to optimisation, quantum key distribution, and new security architectures. They are not priorities for immediate large-scale deployment, but they are lines of development that may shape future competitive advantage.

## Concluding remarks

Banking does not need to compete in the global race to build the largest AI model, but rather in the ability to integrate reliable AI into the core of its financial processes. The advantage will lie in combining efficient models, sovereign infrastructure, regulatory governance, hybrid talent, and a prioritised selection of use cases. The winning AI in regulated sectors will be less spectacular than consumer AI, but more robust, auditable, and defensible.



*AI differentiation in critical sectors will be sovereign by design, efficient by architecture, and regulated from the outset.*

# 01



## Scientific, innovation and talent ecosystem



Seizing the opportunity

The new  
European  
frontier of B2B AI

Strategic Intelligence Report



# Scientific ecosystem

Applied science is shaping financial AI

- » **From an operational universe of 10.3 million AI documents, the analysis isolates 326,000 publications where AI converges with finance and banking.** That core represents only 3.15% of the total, but it concentrates the strategic signal relevant to the sector.
- » **Applied scientific output has surpassed traditional theoretical research.** In 2024, the footprint of applied AI research in OpenAlex exceeds 1.3 million documents, approximately four times more than the foundational literature captured by the established academic benchmarks.
- » **Financial AI is not being built primarily around chatbots, marketing, or low-complexity automation.** The scientific focus is concentrated on high-trust AI, with robust, traceable and explainable algorithms for risk, fraud, creditworthiness and mathematical modelling.
- » **The frontier of innovation no longer appears first in traditional scientific journals.** Preprints act as an early radar for emerging technologies, while journals consolidate validated, auditable knowledge that is closer to banking industrialisation.
- » **The lower visible volume of FinTech and Digital Banking does not point to a loss of relevance, but rather to a privatisation of R&D.** Some critical knowledge is shifting into companies, platforms, and protected corporate silos.

# Definition of the analytical universe

## From the AI universe to the financial core

The analysis reduces AI's overall footprint until it isolates the applicable, measurable and comparable financial signal.

The block starts from **U0\***, the global AI footprint in OpenAlex. The full-corpus measurement reaches **15.6 million documents**, while the consolidated operational corpus totals **10.3 million** for detailed analysis between **2022 and 2025**.

On that basis, specific semantic filtering is applied for banking and finance. **U1** isolates the intersection between AI, finance and banking, with **326,000 documents**, equivalent to **3.15%** of the U0 operational corpus.

The zoom continues with **U2**, focused on sector-specific applications, with **50,748 documents**. **U3** measures quality and impact with a mature citation window, and **U4** reconstructs geographical collaboration networks.

**The methodological value lies in measuring financial adoption, not just academic volume.**

### From the global footprint to the financial core

**U0 · Global AI (operational)**

**10.3M**

OpenAlex 2022-2025



**U1 · AI + Finance**

**326k**

3.15% of total AI  
Semantic filtering against a sector-specific corpus (Banking & Finance)



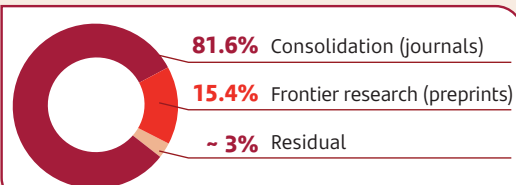
**U2 · Vertical application**

**50.7k**

69.3% Risk + Insurance  
exclude "General Finance"/"Banking"

**Quantitative overview: global AI → AI applied to finance (OpenAlex + AI Index)**

#### Communication channels (U1)



#### Translation for banking: 3 levers

- **Efficiency (Opex)**  
Industrialise Risk/Insurance (mature)
- **Growth (Alpha)**  
Derivatives and markets (high complexity)
- **Trust (compliance)**  
High-Trust AI (governance + validation)

#### Market indicators (CAGR 2022-2024)

- **44.7% Fraud Detection**  
Defensive outburst (urgent)
- **22.7% Derivatives**  
Renewed quantitative focus (investment banking)
- **~ 0% Risk + Insurance**  
Maturity: standard/commodity

#### GenAI ≠ core banking (bibliometric signal)

In specialised literature:  
LSTM/RNN (2,610) > LLM (1,854)  
→ focus on time series + robustness,  
with an emerging semantic layer

\* "U" stands for Analytical Universe, referring to the corpus of scientific documents used in the analysis. U0 captures the overall footprint of papers on AI; U1 filters AI applied to finance and banking; U2 refines the specific vertical applications; and U3 is used to measure quality and impact with a mature citation window.

# AI as an engine of calculation and solvency

## High-trust algorithms in risk and insurance

**Scientific output places financial AI in risk, insurance and regulatory capital above conversational uses.**

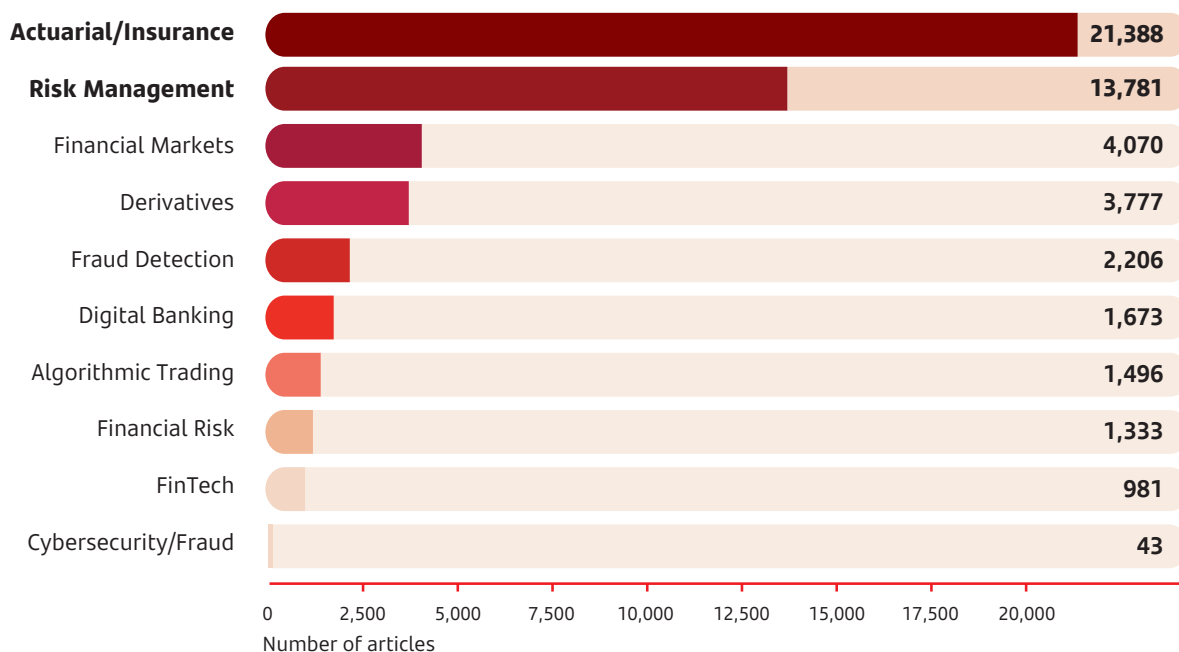
Excluding the general category of banking and finance, the **U2** corpus reveals the true priority of applied research. The **Actuarial/Insurance** category comprises **21,388 articles**, and **Risk Management** accounts for **13,781**.

Both domains account for **69.3%** of specialised research, confirming that AI acts primarily as a numerical engine for capital and solvency calculations. This adoption prioritises the mathematical traceability and auditability required by strict regulations such as **Basel**

or **the European AI Act**. For this reason, references to **explainable AI (XAI)** and the high-trust AI indicator are growing steadily to ensure model risk management and mitigate gender bias.

**Risk management and insurance are adopting AI as an auditable, mature industry standard, far removed from commercial hype.**

### Details of specific AI applications in Finance (excluding general)

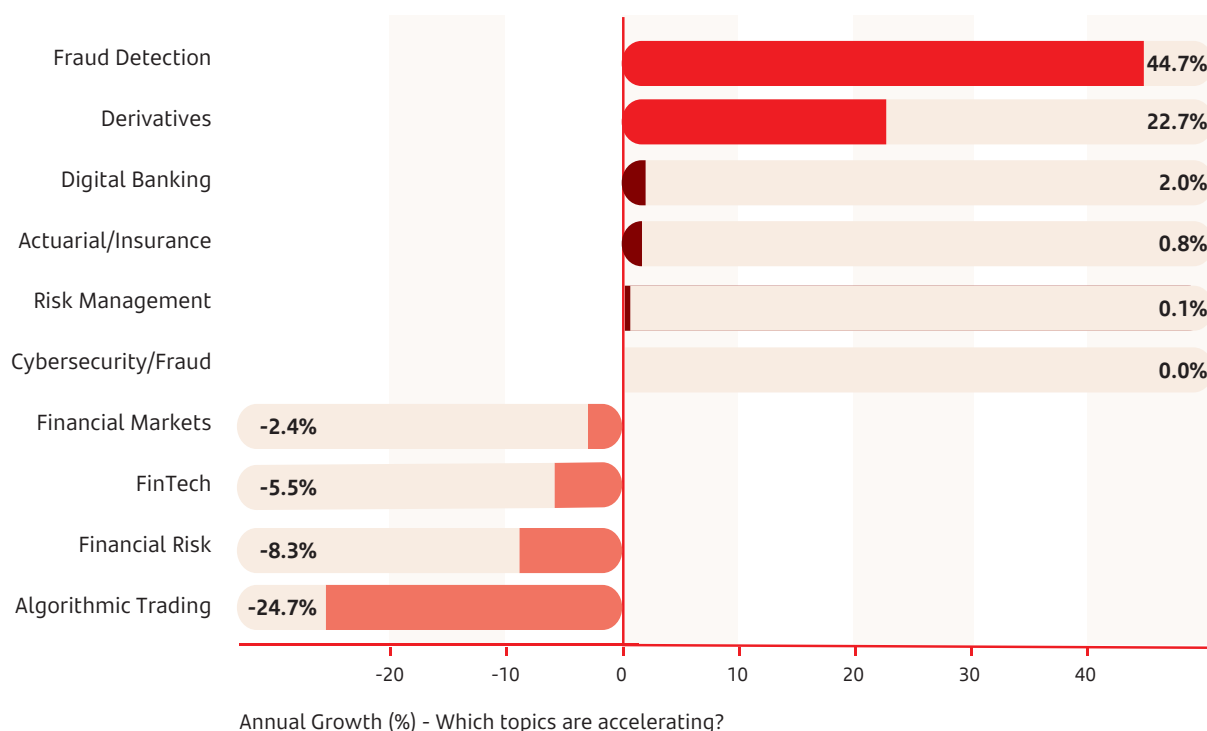




# Financial innovation is advancing at two speeds

The pace of innovation in financial AI is not distributed evenly. Research is accelerating in defensive fields and those of high technical complexity, while some more mature or privatised areas are losing visibility in open scientific output.

## Speed of Innovation: Annual Growth Rate (CAGR 2022-2024)



The upper quadrant (high growth) indicates emerging or returning ("Alpha") trends, while the lower right quadrant (high volume, low growth) identifies the sector's "commodity" or structural technology.

The strong growth in **Fraud Detection** and **Derivatives** points to two clear priorities: security against new threats and quantitative sophistication.

By contrast, the decline of **Algorithmic Trading** and **FinTech** suggests that part of the innovation is shifting towards proprietary environments.

# Two-speed innovation

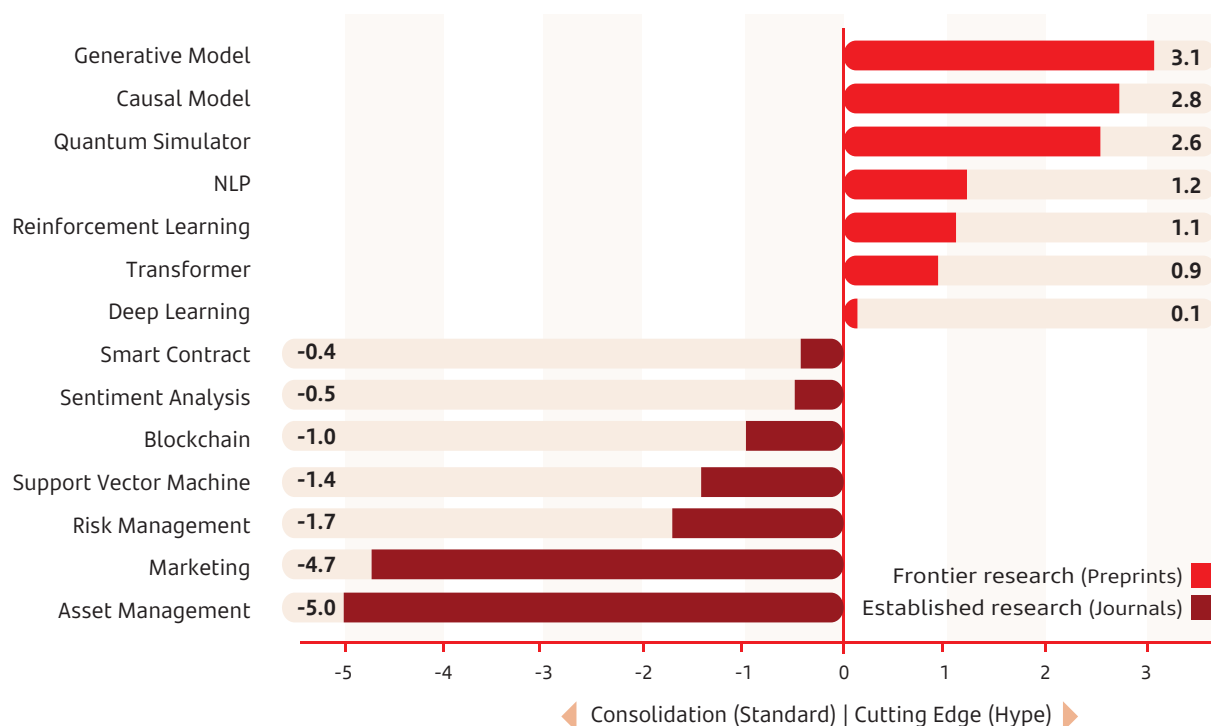
## Scientific validation and early signals

**Financial AI combines auditable knowledge in journals and early signals in preprints, with very different validation times.**

The dissemination of innovation in AI applied to finance shows a **critical asymmetry** in its channels of scientific communication. **81.6%** of knowledge is consolidated in **traditional scientific journals**, which provide the peer validation, transparency and methodological stability required by the banking supervisor for production models.

By contrast, **preprints** account for **15.4%** of the sectoral corpus, far exceeding conference proceedings, which in this particular corpus represent only 3%. This five-to-one ratio suggests that **financial innovation** seeks early authorship, rapid contrast, and less exposure to long editorial cycles.

### Innovation radar: Cutting-edge (Preprints) vs. Established (Journals)



**Overlooking preprints may mean missing early signals, although they should not be taken as definitive evidence.**

# The technology maturity radar

Industrialise today, experiment tomorrow

The divide between journals and preprints separates production-ready technologies from lines still subject to validation.

The radar distinguishes the regulatory readiness zone. **Risk Management, Credit Risk, Blockchain** and **Smart Contracts** appear there, with a greater presence in journals and a clearer use in efficiency, auditing and control.

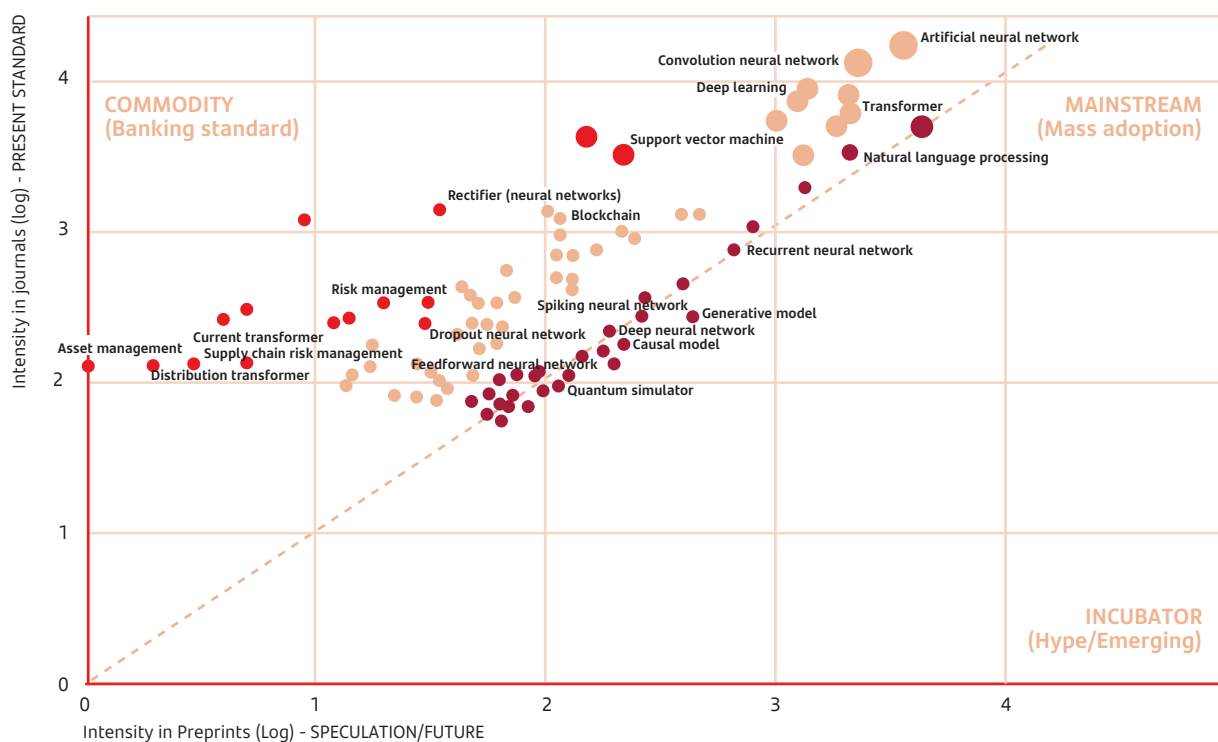
In the emerging area are **generative models, causal AI** and **quantum simulation**.

Their greater weighting in preprints points to technical uncertainty, but also to possible early advantages.

**Deep Learning** occupies an intermediate position. It already operates as cross-cutting infrastructure, connecting prediction, document automation, and decision systems, without replacing deterministic engines in critical processes.

**Bibliometrics determines what to scale, what to monitor, and where to experiment.**

## Maturity Matrix. From the laboratory (Preprints) to Industry (Journals)



Classification of technologies according to their market maturity.

# The geopolitics of knowledge

## Asia gains scientific scale

**Global production reveals regional specialisation, with China dominating risks, emerging Asia leading digital banking while the US retains leadership in frontier innovation.**

The geopolitical map shows that **China** leads in the volume of research applied to finance. Its specialisation is concentrated in **Financial Risk**, accounting for **51.1%** of scientific publications worldwide, and in **Algorithmic Trading**, accounting for **38.3%**.

South-east Asia dominates consumer-oriented banking. **Indonesia** accounts for **48.6%** of research in **Digital Banking** and **30.6%** in **FinTech**.

**India** leads **Fraud Detection** with **43.6%**.

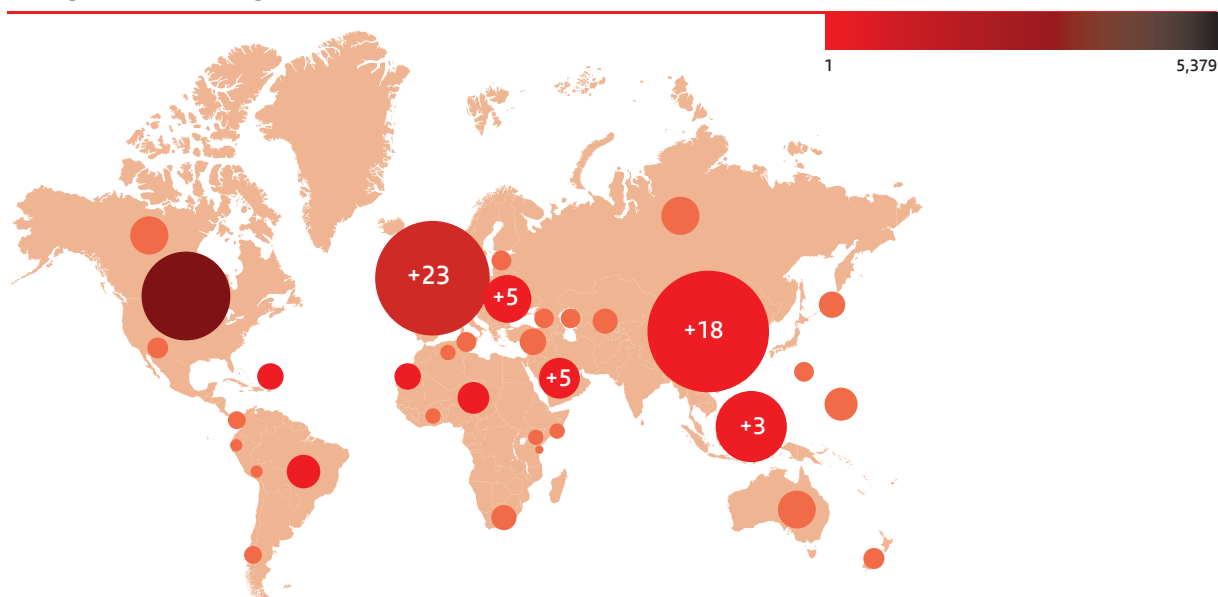
The **US** retains defensive strength in **Cybersecurity** and **Actuarial/Insurance**.

By contrast, the **United States** maintains a dominant position based on the creation of the major frontier foundation models, supported by a hybrid ecosystem with strong industrial investment.

**Europe** lags behind in volume by country, highlighting a concerning fragmentation of research critical mass.

**China dominates financial computation; emerging Asia dominates digital banking. The US dominates frontier foundation models and commercial investment.**

### Geopolitical map



*Ranking in AI applied to Banking and Finance, concentrated by countries and geographical areas.*

# Europe's boutique advantage

## Europe wins on quality

Europe does not lead in terms of volume, but it stands out for quality, specialisation and validation of critical financial technologies.

The European analysis uses **U3**, with **2021** data to avoid recent citation bias. The matrix combines **RCA\***, which measures specialisation, and **Quality Ratio\*\***, which measures impact against the world average.

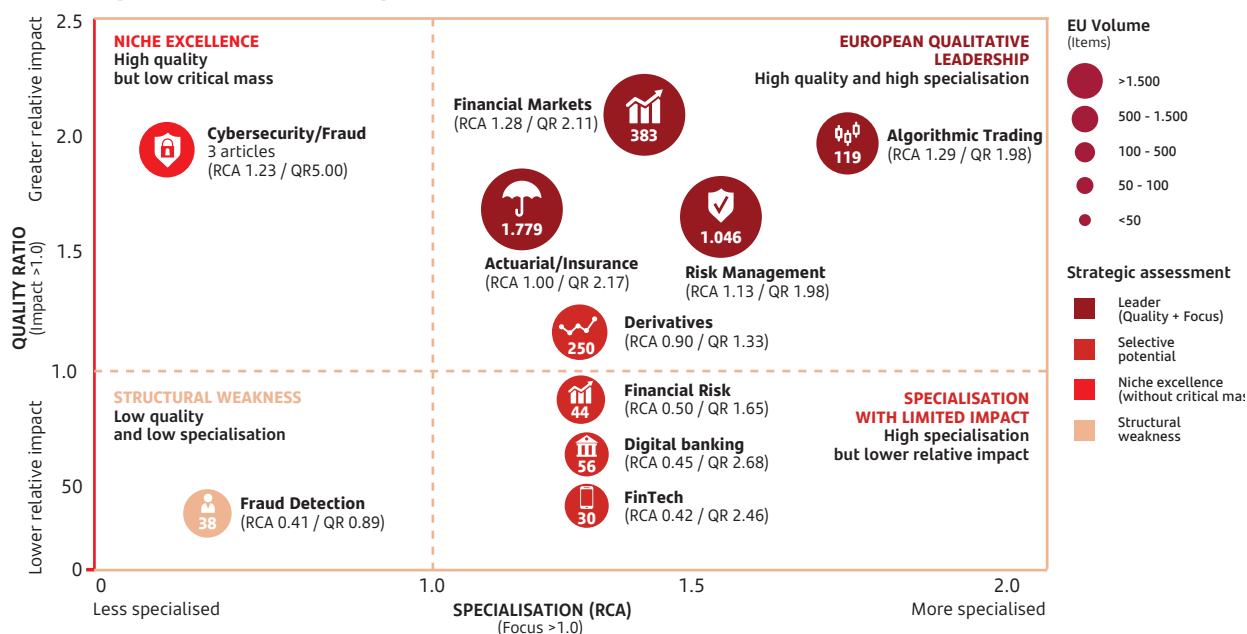
In **Algorithmic Trading**, Europe reaches an **RCA of 1.29** and a **Quality Ratio of 1.98**. In **Financial Markets**, it records an **RCA of 1.28** and a **Quality Ratio of 2.11**, with a strong position in complex markets.

Strength is repeated in **Risk Management** and **Actuarial/Insurance**, with quality ratios of **1.98** and **2.17**. Europe thus positions itself as a global regulatory authority and a technical leader in the validation, certification and supervision of financial systems subject to stringent regulatory requirements.

Europe competes better when quality matters more than scale.

## European competitiveness matrix

Scale, specialisation and impact of research in financial AI. Universe: U3 (Year 2021)



**Interpretation:** Europe does not lead in volume, but it offers clear qualitative advantages in highly complex scientific and financial verticals. The European "Boutique Effect" is not concentrated in markets, trading, risk and insurance.

\* RCA, or revealed comparative advantage, measures whether Europe is more specialised in an area. A value greater than 1 indicates relative specialisation.

\*\* Quality Ratio measures the relative impact of European publications against the world average. A value greater than 1 indicates a higher comparative scientific impact.

# The privatisation of R&D

## Privatisation of science and collaboration gaps

### Digital innovation is moving to proprietary environments.

The low presence of **FinTech** and **Digital Banking** in scientific publications does not imply less innovation. It indicates that a significant part of R&D is shifting towards proprietary companies, platforms and laboratories, where knowledge is protected for competitive advantage and is not always visible in open science.

In **Risk/Insurance**, the industry contributes to **6.6%** of articles, because public validation reinforces supervisory acceptance. In **FinTech/Digital**, industry participation falls to **2%**.

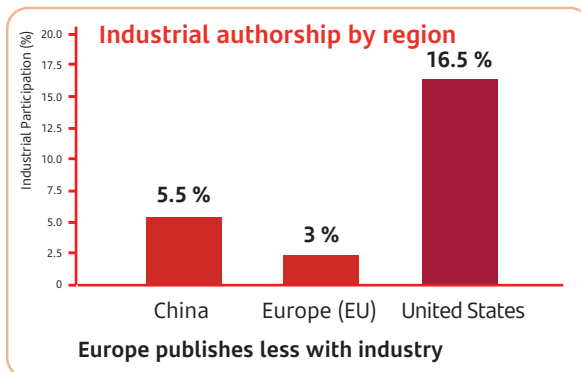
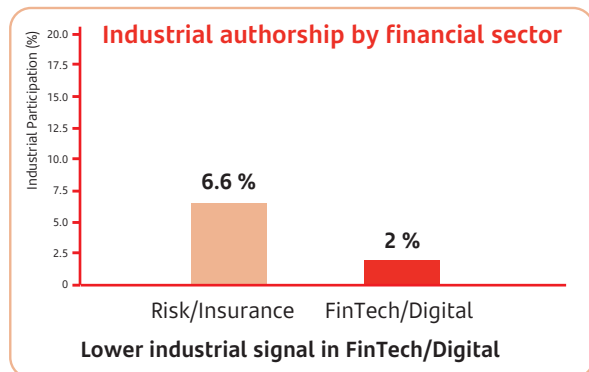
The gap is wider in Europe. Scientific output comes **92%** from academia and only **3%** from industry, compared with the **16.5%** industrial share observed in the United States, which accelerates effective transfer into financial services.

Likewise, cybersecurity appears isolated from solvency, creating a dangerous operational blind spot in the face of systemic requirements such as DORA regulation or operational resilience.

**The lower volume of publications reflects proprietary innovation, not an absence of innovation.**

### Privatisation of science and collaboration gaps

Digital innovation is moving to proprietary environments.



**Interpretation:** Lower industrial authorship does not imply less innovation, but rather a reduced presence of that innovation in open scientific publications.

#### Privatisation of R&D in FinTech.

% of papers with business/industrial affiliation (2025)  
FinTech/Digital shows a weaker industrial signal than Risk/Insurance.

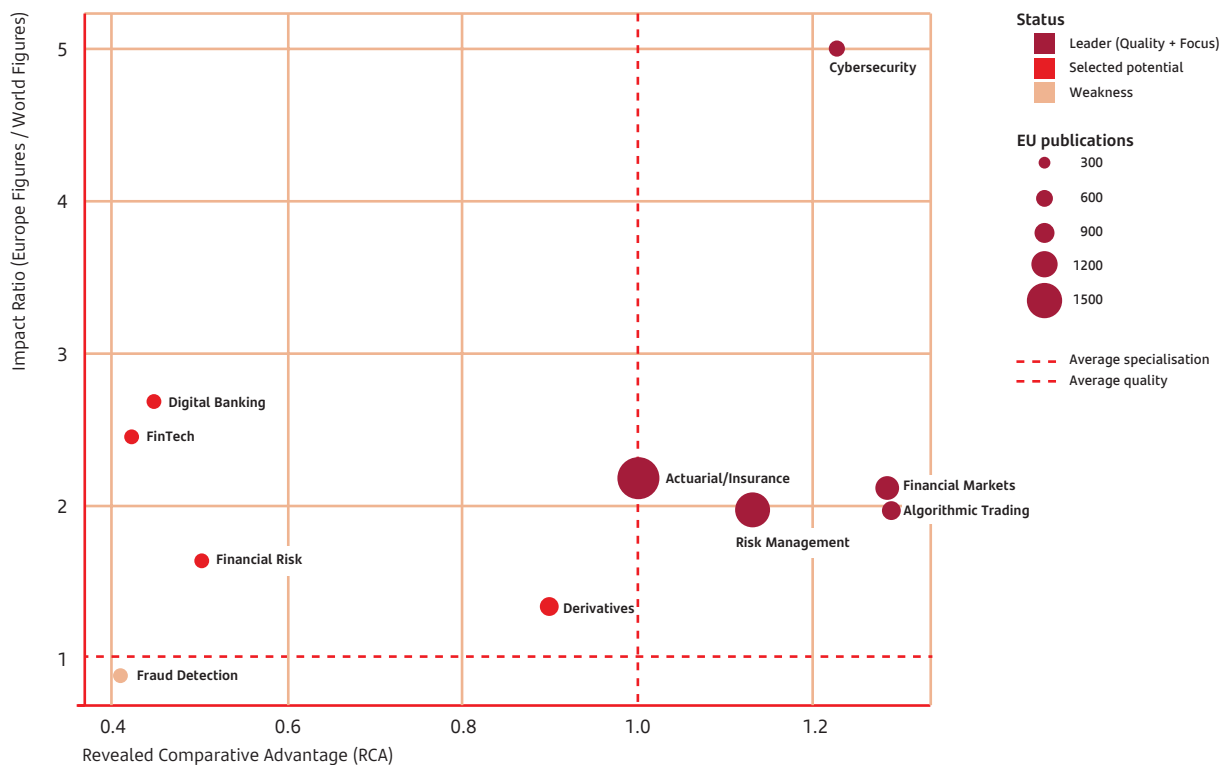
#### Institutional Asymmetry.

Composition of the scientific authorship.  
Source: Authors' analysis based on OpenAlex and the AI Index Report 2025

# European cybersecurity, non-scalable excellence

European research into financial cybersecurity presents a notable anomaly. The **Cybersecurity/Fraud** vector reaches the highest relative impact in the matrix, but does so with minimal output, revealing highly concentrated excellence that is difficult to scale.

## European Competitiveness Map 2021: Specialisation vs. Real Quality



The dark blue point can be seen in the top right-hand corner (Cybersecurity). Its isolated position reveals concentrated excellence, but it is difficult to scale operationally.

The result points to a systemic vulnerability. Europe has research groups capable of producing world-leading science, but the lack of critical mass limits their ability to cover the full scope of financial, operational and digital threats.

# The national competitiveness map

## Spain develops applied science

Spain combines critical mass and scientific output in universities, regulatory specialisation, Latin American links, and a **CSIC** focused on **symbolic and logical AI**.

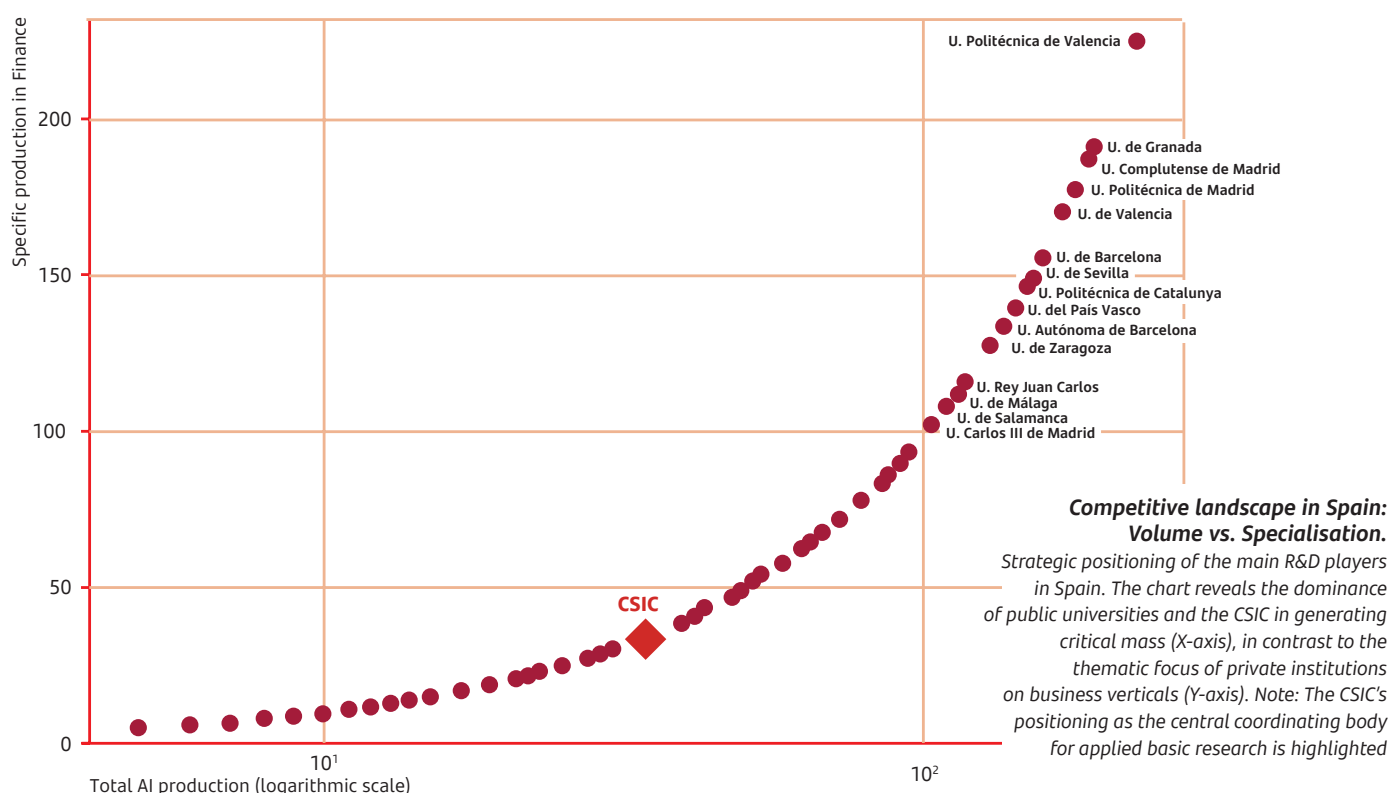
Public universities clearly lead in generating critical mass and overall volume of scientific output, focusing their specialisation on machine learning and advanced deep learning models. The national map places **UPV, UGR** and **UCM** among the driving forces of production in financial AI. Spain shows strong penetration in hybrid spaces between technology, economics, regulation and markets.

The thematic analysis identifies a significant emphasis on **law, regulation** and market microstructure. This signals a natural position in **RegTech**, algorithmic supervision and the legal interpretation of financial AI.

By contrast, the **CSIC** serves as the backbone of **symbolic AI**, the design of multi-agent systems, and formal logical reasoning. This methodological specialisation is essential for financial regulatory compliance, providing the mathematical interpretability, ethics and auditability required to certify high-risk models under the new EU regulation.

**The university provides scale; CSIC provides logic, interpretability and technical depth.**

### Competitive landscape in Spain: Volume vs. Specialisation





# Innovation ecosystem

European supernodes of excellence

- » **Europe does not operate as a homogeneous AI ecosystem, but rather as a polycentric network of specialised nodes.** Each hub or supernode brings a different capability to the financial sector: algorithmic efficiency, industrial trust, agentic reasoning or regulatory validation.
- » **The analysis changes scale.** After measuring scientific output, the focus shifts to technology transfer, spin-offs, frontier laboratories and B2B suppliers capable of turning knowledge into a regulated B2B product.
- » **Paris** offers **frugal models and sovereignty** of execution. The **German-Swiss** (DACH) hub connects **industrial AI**, explainability and critical safety. **London** leads in advanced reasoning and **agentic AI**. **Spain** strengthens **sovereign computing, supervision and trust**.
- » **University spin-offs are the bridge between scientific excellence and competitive advantage.** They enable the retention of intellectual property, the acceleration of pilot projects and new use cases, the attraction of specialised talent, and the testing of solutions before the market consolidates them.
- » **For a financial institution, this radar is not an academic mapping exercise, but a decision-making tool** for prioritising partnerships, pilot projects and strategic monitoring in the areas where European AI can deliver sovereignty, efficiency and compliance.

# Paris Node

## The mathematical epicentre and model architecture

**The French node combines mathematical tradition, frugal models and open weights to reduce technological dependency and inference costs.**

**Paris** is consolidating its position as the European epicentre of algorithmic efficiency. Its advantage lies not in competing through computational brute force, but in decoupling performance and size through **SLMs**, **MoE** architectures, and advanced data curation.

**MoE** models, or mixture of experts, activate only a specialised part of the model for each task. This makes it possible to improve efficiency, inference cost and performance without linearly increasing computational consumption.

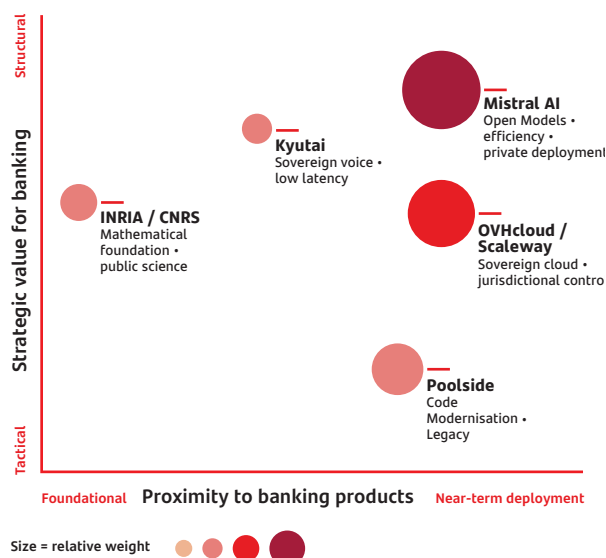
**Mistral AI** represents the enterprise layer of the node, with open-weight models and controlled deployment options in private environments. For regulated sectors such as banking, this enables data sovereignty, reduced exposure to external clouds, and operational control over sensitive models.

**Kyutai**, together with **INRIA** and the **CNRS**, strengthens the open science, sovereign voice, biometric privacy and mathematical transfer dimensions towards tightly controlled banking applications. SODA validates **Scikit-learn** in fraud and risk with partners such as **AXA** and **Allianz**.

### Paris Node

#### Node capability map

**Paris** is coordinating a sovereign AI ecosystem supported by public science, frugal models and private deployment for regulated environments.



**Interpretation:** qualitative positioning of the Paris ecosystem for systemic banking.

#### Model keys



The differentiating advantage lies in algorithmic efficiency.



Open-weight models reduce technological dependence.



The sovereign cloud strengthens control and privacy.

**Paris allows banking AI to scale without mortgaging sovereignty, operating margin or banking secrecy.**

# Munich node and DACH axis

## Industrial AI, sovereignty and technical trust

Germany and Switzerland contribute industrial AI, explainability, critical security and post-quantum technologies applicable to corporate banking.

The **Munich-Zurich-Stuttgart** axis acts as the major European hub for industrial AI. Its focus is B2B, traceable and robust, with the ability to bring models into physical processes, supply chains and financeable productive assets.

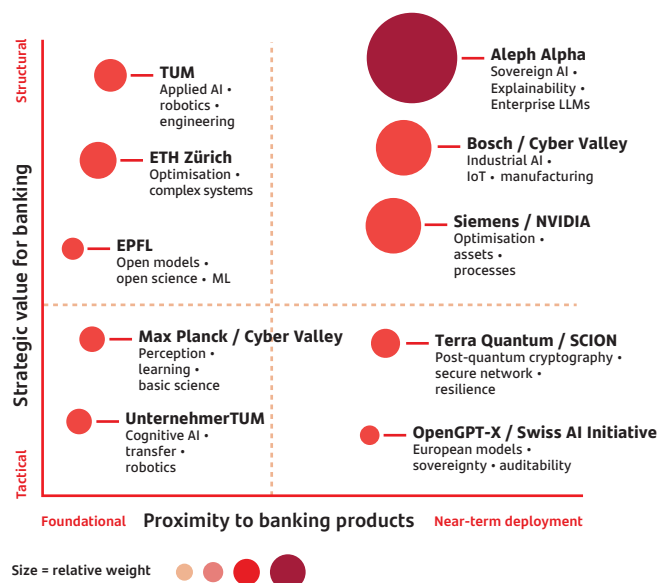
The first layer is scientific. **TUM**, **ETH Zurich**, **EPFL** and **Max Planck** contribute excellence in causality, graphs, robust learning and reliable systems. The second layer is transfer, with **UnternehmerTUM** and **Cyber Valley** acting as bridges between advanced research, industry and technological entrepreneurship.

The third layer reinforces sovereignty and trust. **Aleph Alpha**, **OpenGPT-X** and **the Swiss AI Initiative** drive explainable and controllable models, while **Terra Quantum** adds resilience, critical security and post-quantum capabilities. For banks, the differentiating value lies in turning physical assets and industrial processes into verifiable data for corporate lending, insurance and payments.

### DACH node

#### Node capability map

**DACH** turns AI into trusted infrastructure for banking and industry, combining explainability, critical security, and applied transfer to the productive sector.



**Interpretation:** qualitative positioning of the DACH ecosystem for systemic banking.

#### Node keys



The differentiating strength lies in reliable industrial AI.



Explainability and sovereignty strengthen control.



Post-quantum security broadens banking resilience.

**DACH makes it possible to assess industrial assets using operational data, not just historical balance sheets.**

# London node

## The convergence between Bio-Fintech and agentic AI

The British node focuses on advanced reasoning, multi-agent systems, financial graphs and commercial validation in B2B fintech.

**London** combines the City, **Google DeepMind**, the **Alan Turing Institute** and the **Oxford-UCL-Cambridge** triangle. Its advantage lies not only in generating models, but in applying the science of complex systems to markets, risk, compliance and financial decision-making.

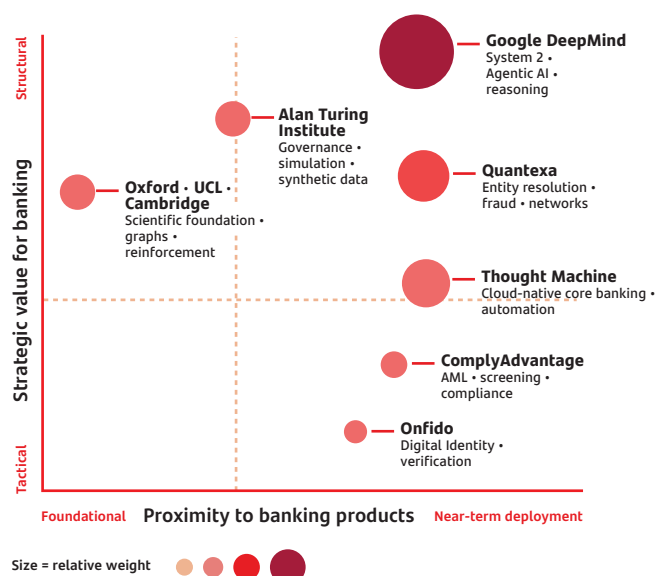
The node brings together capabilities in **system 2**, reinforcement learning, graph networks, synthetic data, and **agentic AI**. DeepMind is accelerating this agenda with approaches such as **Finance-Agent**, while Isomorphic Labs is extending bio-AI into the modelling of complex systemic risks.

The **Alan Turing Institute** contributes ethical governance, market simulation and synthetic data alongside actors such as **HSBC** and the **FCA**. In parallel, B2B fintechs such as **Thought Machine**, **Quantexa**, **OakNorth**, **Onfido** and **ComplyAdvantage** validate these architectures in real financial processes.

### London node

#### Node capability map

**London** combines System 2 science, agentic AI and B2B fintech to apply complex models to risk, compliance and banking modernisation.



**Interpretation:** qualitative positioning of the London ecosystem for systemic banking.

#### Node keys



The differential advantage lies in advanced reasoning.



B2B fintech accelerates the financial application.



Governance and synthetic data strengthen control.

**London transforms analytical AI into AI for controlled execution.**

# Spain node

## The pillar of infrastructure and trust

**Spain provides sovereign infrastructure, early oversight and applied ethics to scale banking AI under legal certainty.**

**Spain** is not competing primarily in large commercial models. Its distinctive position lies in sovereign computing, regulatory validation, and trust. **MareNostrum 5**, the **BSC-CNS** and the **AI Factory** offer public capacity to train and audit sensitive models. **AESIA** provides a unique regulatory asset. The sandbox enablaes the testing, fraud or automated decision-making models before deployment, reducing contingencies under the **AI Act**.

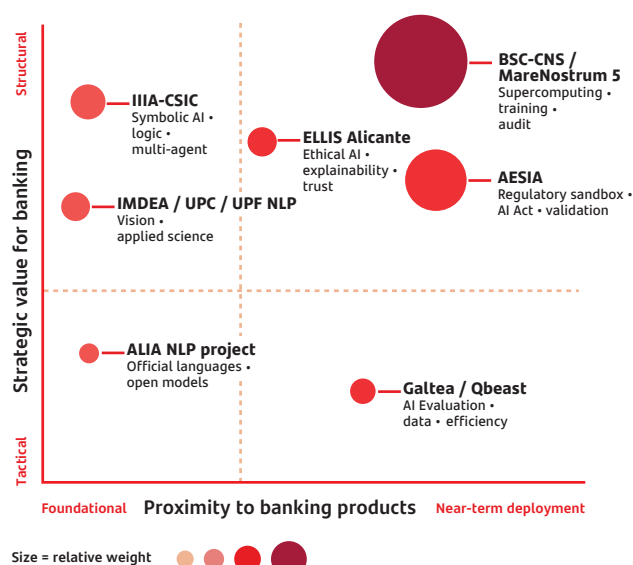
The **ELLIS Alicante** network, together with **IMDEA**, **UPC**, **UPF** and **IIIA-CSIC**, completes the node with algorithmic ethics, privacy, explainable AI, NLP and autonomous systems.

The spin-offs **Galtea** and **Qbeast** connect this scientific foundation with tools applicable to model evaluation, bias control, and efficiency in data management.

### Spain node

#### Node capability map

**Spain** is building a trustworthy AI ecosystem based on sovereign supercomputing, regulatory validation and open research networks for highly controlled banking environments.



**Interpretation:** qualitative positioning of the Spanish ecosystem for systemic banking.

#### Node keys



Sovereign infrastructure accelerates training and auditing.



Regulatory validation reduces deployment risk.



Ethical and symbolic AI strengthens trust.

**Spain could be the European testing environment for high-risk financial AI.**

# The leap from the laboratory to the market

## From radar to regulated B2B adoption

**The European ecosystem of start-ups and spin-offs transforms research, data and infrastructure into capabilities applicable to financial services and regulated sectors.**

European AI innovation is not concentrated in a single technology or a single player. The radar shows an ecosystem composed of complementary capabilities that address specific challenges in the financial sector: regulatory compliance, fraud prevention, risk management, document automation, data privacy, technological modernisation and digital sovereignty.

To convert this technology map into operational decisions, the analysis is organised using a prioritisation matrix that cross-references the banking adoption horizon with the structural impact on the entity's architecture. This interpretation makes it possible to distinguish solutions for immediate application, scalable capabilities through collaboration, medium-term strategic bets, and technologies that require monitoring before reaching sufficient maturity.

The areas closest to adoption are concentrated in RegTech, AML, KYC, fraud and document intelligence, where solutions already exist that can reduce operational costs, speed up processes and strengthen regulatory compliance.

Beyond operational efficiency, the radar identifies a second layer of enabling technologies focused on privacy, synthetic data, federated learning, and model governance. These capabilities make it possible to develop AI based on sensitive information without compromising confidentiality, banking secrecy or regulatory requirements.

The documentary dimension constitutes another of the main return vectors. Contracts, case files, regulatory reports and financial documentation continue to account for a significant part of organisations' operational workload. New document processing, semantic graph and knowledge reasoning solutions make it possible to transform unstructured information into governable and reusable assets.

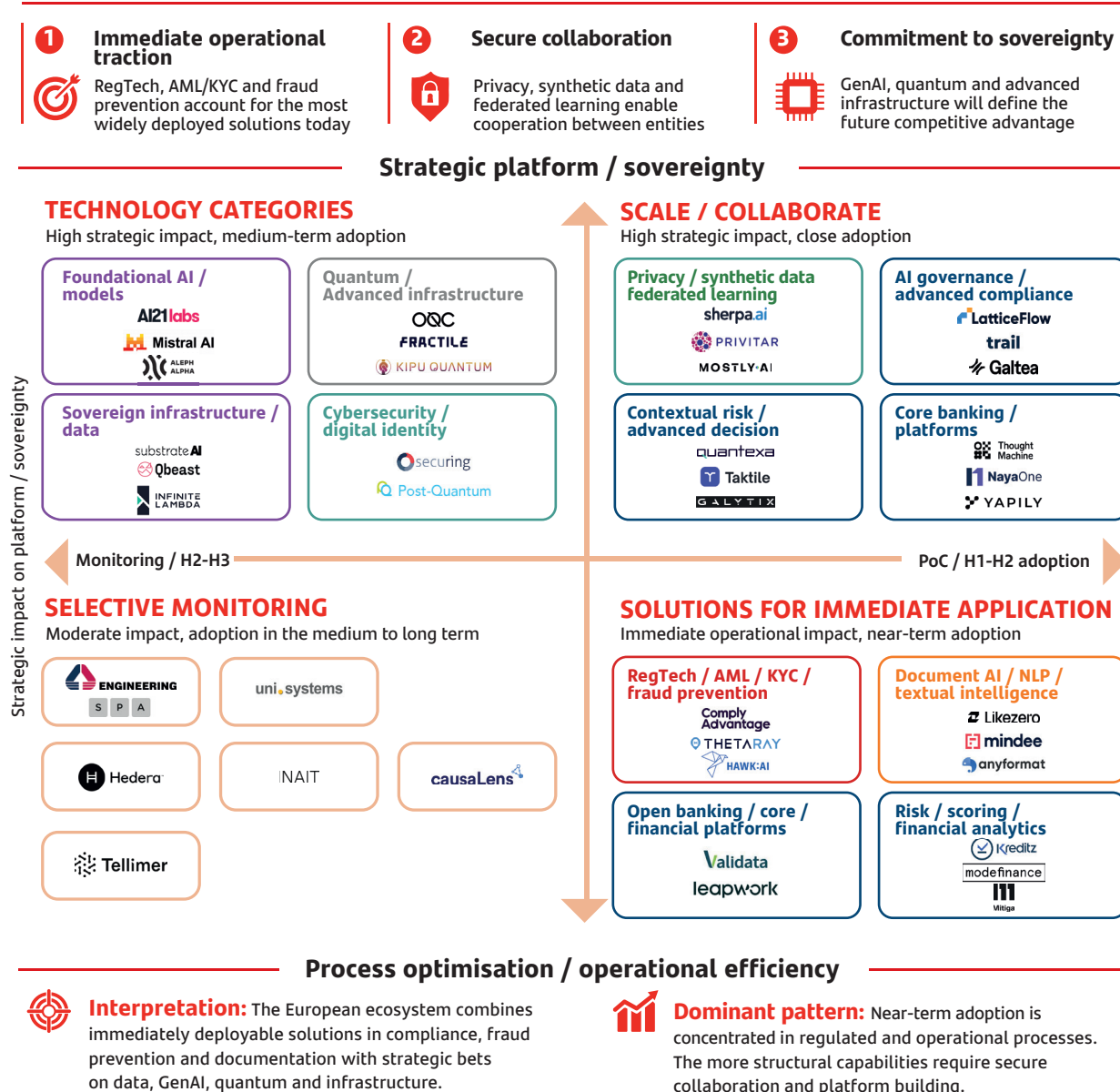
Finally, the ecosystem incorporates broader-reaching ventures linked to foundation models, advanced infrastructure, cyber security, quantum computing and next-generation financial platforms. Although its adoption will be gradual, these capabilities could redefine the technological architecture of banking over the next decade.

**The competitive advantage will not come from incorporating more AI, but from deploying governed, secure AI aligned with the entity's regulatory architecture.**

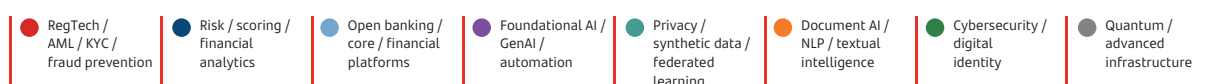
# European B2B AI Radar for Financial Services

Technological maturity, adoption horizon and strategic impact for regulated companies

## Maturity matrix, strategic impact and technological capabilities of the innovation ecosystem



### Technology categories



**NOTE:** Selected companies representative of the European AI ecosystem for financial services.  
Horizons: H1 = 0–12 months | H2 = 12–24 months | H3 = 24–36 months | PoC = Proof of concept

# Transfer, talent and governance

From external innovation to internal capability

- » The advantage lies not only in identifying science, start-ups, spin-offs or European nodes of excellence. **The differentiating factor is building an internal architecture capable of detecting and absorbing that knowledge, turning it into useful pilots and scaling it in a controlled way within a regulated entity.**
- » **Lab to Market transfer requires choosing the appropriate vehicle according to the challenge.** Industrial co-location, scientific venture building, sovereign spin-offs, licences, B2B consortia and co-investment do not address the same problem, nor do they generate the same type of advantage.
- » **The battle for talent is not resolved solely by hiring technical profiles on the market.** Banking needs to combine external recruitment, internal training, scientific partnerships and academic retention to create hybrid profiles capable of operating AI in complex environments.
- » **The critical profile is the professional who translates between business, data, technology and control.** Without the Analytics Translator profile, integrated teams and specific governance roles, AI may work in pilot mode, but become blocked when moving into production.
- » **The AI Act turns governance into an operational capability.** The entity shall classify systems, assign responsibilities, document models, validate changes, monitor behaviour and retain evidence throughout the life cycle.



# Market transfer models

## Transferring science requires architecture

The advantage lies not only in identifying innovation, but in absorbing it through stable vehicles for transfer, intellectual property and B2B scaling.

Europe has developed various frameworks for turning science, talent and intellectual property into business solutions. A first avenue is **corporate co-location ecosystems**, that is, environments where universities, scientific centres, industrial laboratories and large companies work closely together from the early stages **Cyber Valley** and **UnternehmerTUM** are two examples of this model.

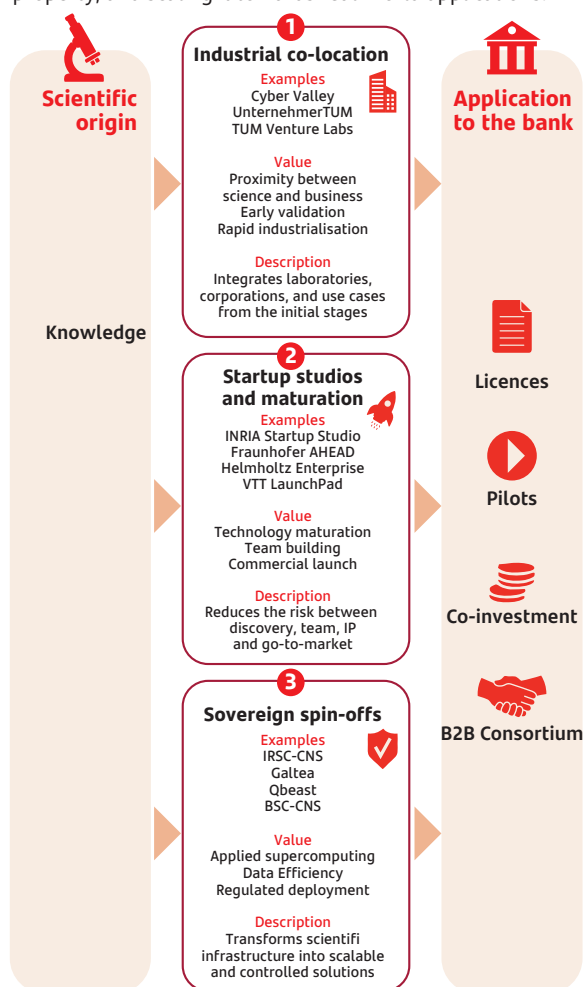
A second route is **scientific maturation and venture-building programmes**. They are not a specific company, but rather instruments created by research centres to transform a promising technology into an investable company, licence or product. **INRIA Startup Studio**, **Fraunhofer AHEAD**, **Helmholtz Enterprise** and **VTT LaunchPad** operate according to this logic, reducing risk between discovery, team, intellectual property and commercial launch.

A third route is **spin-offs born from sovereign scientific infrastructures**, such as the **BSC-CNS**, with cases such as **Galtea** or **Qbeast**. For banks, the key is to choose the right channel according to the challenge, whether that means exploring frontier AI, licensing mature technology, co-investing, or participating in federated learning consortia.

### Transfer models

#### Three ways to capture science, talent and IP

Europe turns research into business solutions through stable transfer architectures. For banking, the challenge is not detecting innovation, but absorbing it, protecting its intellectual property, and scaling it towards real-world applications.



**Interpretation:** effective transfer does not buy innovation; it builds channels to absorb it.

**Effective transfer does not buy innovation; it builds channels to absorb it.**

#### Model keys

- Each challenge requires a different transfer vehicle.
- IP and talent must be secured before scaling.
- Banking value arises from connecting science with regulated deployment.

# The battle for talent

## Talent limits scale

**The shortage of technical and hybrid profiles makes AI industrialisation a structural constraint for the banking sector.**

Spain presents a significant paradox. There is a relatively solid general-purpose digital foundation, but the advanced critical mass remains insufficient to deploy **AI**, data, architecture and operations in complex environments.

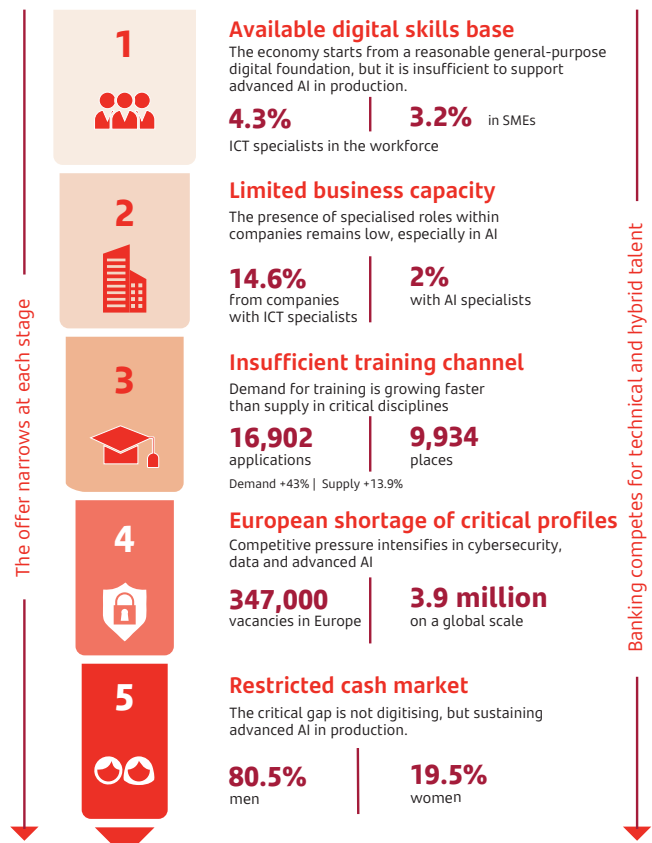
The gap becomes visible within companies. There is a low presence of ICT specialists and an even lower penetration of AI specialists, which limits the shift from pilot trials to stable capabilities.

The problem is not solved simply by hiring externally. European competition for the same profiles, the territorial concentration of talent and the brain drain towards Big Tech make it necessary to combine recruitment, retention, internal training and collaboration with scientific networks.

**The critical gap is not digitising, but sustaining advanced AI in production.**

### Talent constraints funnel

**How the effective supply of ICT and AI profiles is narrowing for banking and industry**



#### Result

**Difficulty in industrialising advanced AI in banking.**  
The constraint is not only attracting talent, but sustaining it, combining it, and scaling it

#### Keys



The shortage starts not with recruitment, but with the available talent pool.



Advanced training is not growing at the pace of business demand.



Without hybrid profiles, AI remains at the pilot stage and does not scale to production.

# The critical profile: Analytics Translator

The Analytics Translator turns AI into business value

**The Analytics Translator\* profile connects business, technology and control so that AI use cases are viable, measurable and deployable.**

In the industrialisation of banking AI, the most critical profile is not always the most specialised data scientist, but rather the professional capable of translating between areas that normally work with different languages. This role is commonly known as an **Analytics Translator or operational value translator**.

Their role is to connect the needs of the business with the real possibilities of the data, the models, and the technological architecture. In banking, moreover, **risk, compliance, audit, privacy** and **traceability** constraints must be incorporated from the outset, because an AI solution cannot be assessed solely on its technical accuracy.

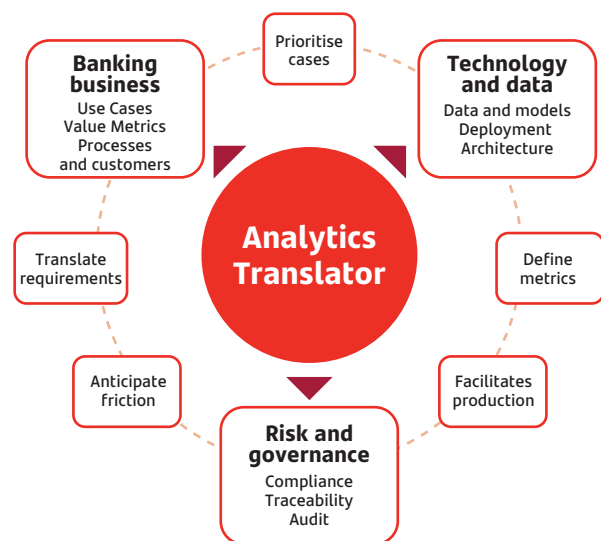
That is why a **business expert with advanced AI skills** can generate more immediate value than a technical profile without regulatory context. Their contribution lies in prioritising use cases, defining metrics, anticipating friction and turning a functional proof into a governed, operable and defensible solution.

**Without operational translation, AI remains at the pilot stage and fails in production.**

## Analytics Translator

### Bridge profile between three domains

They connect business, technology and control to turn an AI use case into a viable, deployable and governed banking solution.



**Interpretation:** the value of the profile lies in translating operational needs, analytical capabilities and control requirements into deployable decisions.

### Keys



The role does not replace business or data teams; it connects the two.



They integrate value, viability and regulatory requirements.



They stop AI from staying on autopilot and accelerate its move into production.

\* Analytics Translator is the bridging role that turns business needs into viable, measurable and governable AI use cases.

# From the functional silo to the AI team

## Integrated teams to industrialise AI

**Pilot projects do not scale when business, technology, risk, and compliance become involved sequentially. Production requires shared teams from the design stage.**

The traditional organisational model separates functions by specialism. The business defines the need, technology builds the solution, risk and compliance review it afterwards, and operations receives the system at the end. This sequence may work in conventional projects, but in AI it often leads to bottlenecks, rework, and pilots that never reach production.

The alternative is **multidisciplinary AI teams**, also known as **Fusion Teams**. They are not coordination committees, but rather working groups with business, data, architecture, risk, compliance, security, privacy and operations profiles participating from the outset of the case.

Their value lies in anticipating the real deployment conditions. The team not only asks whether the model works, but whether it can be integrated with the available data, operate with traceability, pass independent validation, be maintained in production, and generate sufficient evidence for audit and supervision.

**AI does not scale when each department validates too late what it should have designed from the outset.**

### AI requires integrated teams

From the sequential siloed workflow to the Fusion Team capable of taking an AI use case through to governed production.

#### Siloed sequential model

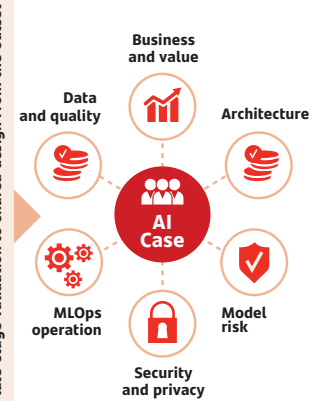


**Result:**  
Duplicated work, bottlenecks and pilots that do not scale



**Interpretation:** AI does not scale when each area validates too late what it should have designed from the outset.

#### AI Fusion Team



**Result:**  
Viable design, controls from the outset and move into production

#### Keys

1. Anticipate risks before deployment.
2. It integrates data, architecture and control.
3. It turns pilots into operational capabilities.

# Governing AI requires new roles

## Roles for industrialising AI

**AI governance must be designed before deployment, not added at the end.**

**Banking requires specific functions to integrate architecture, risk, compliance and evidence throughout the AI lifecycle.**

When an AI use case moves from pilot to production, it stops being merely a technological development. It must be integrated into real-world systems, use governed data, comply with regulatory requirements, withstand audits, and remain under supervision throughout its lifecycle.

That is why new governance roles are needed. The **enterprise AI architect** ensures that the solution fits with the bank's architecture, legacy systems, security, and resilience

and day-to-day operations. Its function is to prevent a good prototype from becoming a solution that is impossible to deploy or maintain.

**AI governance** profiles translate risk, compliance, validation and auditability into verifiable controls. Together with the data owner, independent validation, security, privacy and MLOps, it turns AI into a governed operational capability, rather than a collection of isolated pilots.

## Roles for industrialising AI

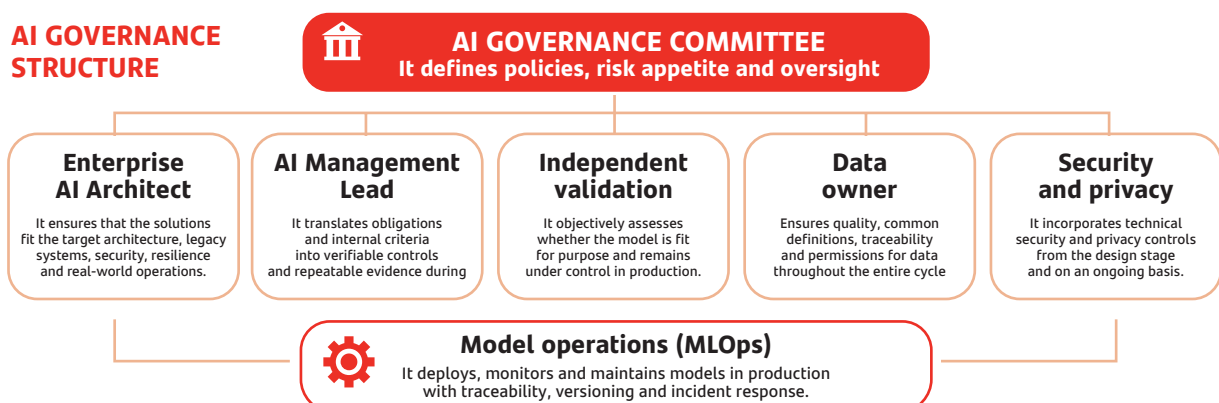
**AI does not scale with data scientists alone. It needs functions that connect architecture, validation, compliance, operations and evidence.**

When an AI use case moves from pilot to production, it stops being merely a technological development. It must be integrated into real-world systems, use governed data, comply with regulatory requirements, withstand audits, and remain under supervision throughout its lifecycle.

That is why new governance roles are needed. Each function contributes a critical piece to ensure the solution is deployable, controllable and defensible over time. Without these functions, the entity accumulates pilots, but not governed capabilities.

**“ KEY IDEA** Governing AI is not about documenting at the end, but about designing controls from the outset **”**

### AI GOVERNANCE STRUCTURE



**Common objective:** To turn AI into an operational, governed, secure and sustainable capability that generates value and complies with regulatory requirements.

### Shared responsibilities throughout the life cycle

**Design**  
Definition of the case, data, model, metrics and usage limits

**Development and testing**  
Model construction, technical validation and quality testing

**Validation and approval**  
Independent review, model risk and compliance

**Deployment**  
Integration, pre-checks and transition to production

**Operation and monitoring**  
Continuous supervision, revalidation, auditing and continuous improvement

# Organisational hybridisation

## Train in three layers

**Upskilling should turn business experts into hybrid profiles through pathways differentiated by responsibility, risk, and technical capability.**

Organisational hybridisation does not consist of training the entire workforce to the same technical level. A financial institution needs phased training, in which each group understands AI according to its exposure to risk, its decision-making capacity and its role in deployment.

The first layer creates a common foundation of **safe literacy**. It should help people use AI tools with discernment, understand their limits, protect sensitive information, and recognise risks related to privacy, bias, security, or misuse. It is the minimum requirement for widespread adoption without increasing operational risk.

The second layer develops **hybrid profiles** in the areas of business, risk, compliance, operations, payments, finance or product. Here, professionals are trained who are capable of translating real needs into use cases, working with data teams, and assessing whether a solution can be deployed with sufficient controls.

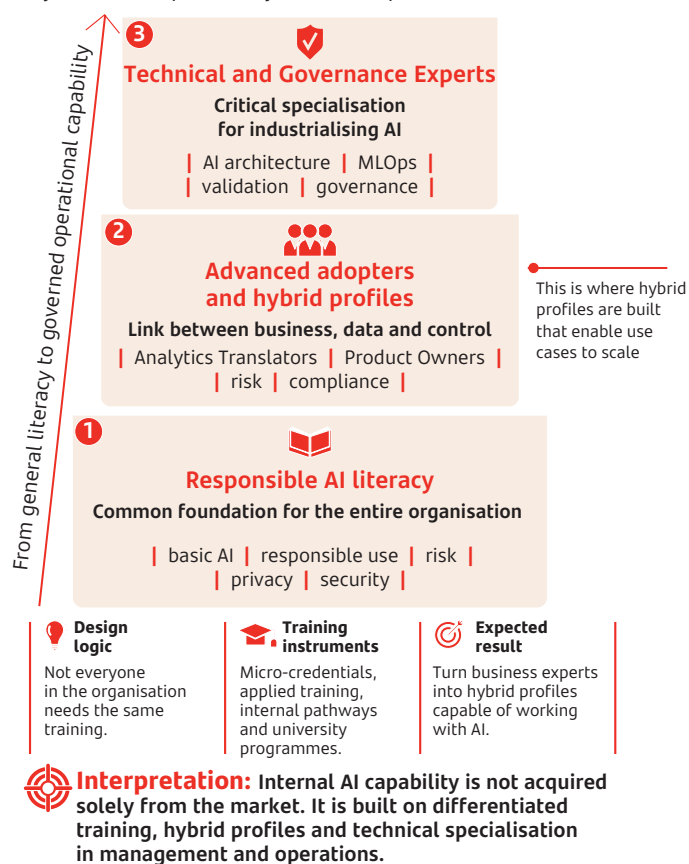
The third layer is reserved for advanced specialists in **AI architecture, MLOps, validation, privacy, security** and **governance**. Its role is to ensure that models not only work, but can also be maintained, audited and evolved throughout their lifecycle.

**Internal capability is built in layers, not through mass generic training.**

**Micro-credentials**, internal pathways and applied university programmes make it possible to accelerate this transition through short, certifiable training linked to the bank's real-world challenges.

### Train in three layers

Internal AI capability is built through pathways differentiated by level of responsibility, risk, and specialisation.



#### Keys

- Safe literacy creates a common foundation.
- Hybrid profiles connect business and technology.
- Advanced experts ensure deployment and governance.

# Academic retention instruments

Retaining scientific talent cannot be resolved with a single lever

**Industrial doctorates, chairs and networks of excellence make it possible to retain talent and capture applied intellectual property.**

The response to the **talent shortage** cannot depend solely on the labour market. Instruments such as **industrial doctorates**, **ENIA chairs** or **Chip chairs**, the **ELLIS** networks and partnerships with centres such as the **CSIC** make it possible to anchor scientific knowledge to specific banking challenges.

The industrial doctorate transforms a complex problem into applied research with academic and business supervision.

The chairs offer continuity, reputation and a shared agenda in responsible AI, data, cybersecurity or hardware.

For banks, these instruments serve a dual function. They reduce dependence on outsourcing and create a stable pipeline of profiles with scientific knowledge, operational context and the potential to generate intellectual property.

## Retention instruments matrix

### Training, applied research and intellectual property acquisition to retain scientific talent

Retention no longer depends solely on salary. It requires combining training, applied research, public funding, scientific networks and the capture of intellectual property

| Instrument                                                                                                               | Main use                                           | Horizon             | Commitment  |
|--------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------|---------------------|-------------|
|  <b>Microcredentials / Upskilling</b> | Retrain functional profiles into hybrid profiles   | 6-12 months         | Low         |
|  <b>Industrial doctorates</b>         | Solve complex problems with applied R&D            | 3-4 years           | Medium      |
|  <b>ENIA Chairs / Chip</b>            | Create stable AI, data or hardware lines           | 2-4 years           | Medium-high |
|  <b>Choose Europe / SpAIn Hub</b>     | Attract senior international or repatriated talent | Medium to long term | Medium      |
|  <b>ELLIS Networks / Max Planck</b>   | Access top-tier researchers and frontier science   | Continuous          | High        |
|  <b>Spin-offs / KIBS</b>              | Capture deep tech, IP and applied transfer         | 3-5+ years          | Variable    |



**Interpretation:** The winning strategy combines short-term instruments to retrain internal talent with medium- and long-term vehicles to capture science, IP and differentiated capabilities.

*A guiding reference for prioritizing academic retention mechanisms and technology transfer in banking.*

**The best retention does not compete on salary; it offers mission, science and a career path.**



# Internal regulatory alignment

## Preparing to audit algorithms

**The AI Act requires governance, documentation, validation and oversight to be turned into internal operational capabilities.**

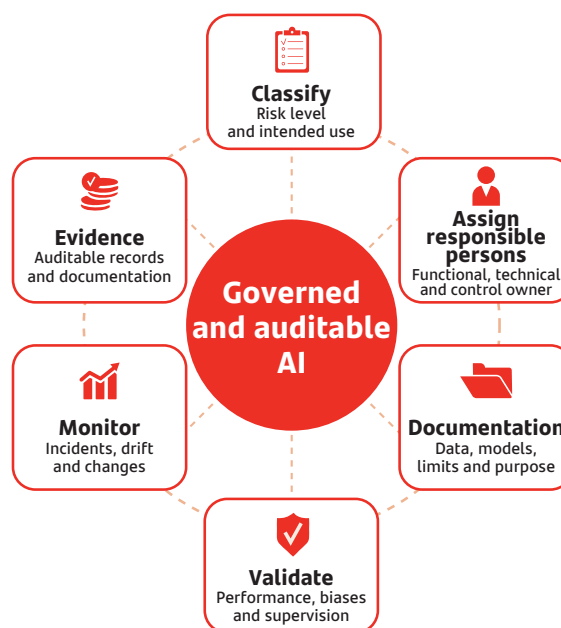
European regulation changes the way a financial institution must deploy AI. It is no longer enough to approve a use case from the legal department or document the model at the end of the project. In higher-risk systems, the entity will have to demonstrate that the model has been designed, tested, monitored and maintained with verifiable controls.

This turns compliance into an operational and technical task. Each model must be capable of being classified, linked to a responsible person, recording its training or usage data, justifying its changes, and retaining evidence relating to performance, bias, human oversight and incident management. The question will no longer be only whether the model complies, but whether the entity can demonstrate it throughout its entire life cycle.

To achieve this, the organisation must connect **AI governance**, model risk, data protection, security, internal audit, and business stakeholders. The objective is to move from general policies to **genuine system traceability**, where each relevant decision can be explained, reviewed, and upheld under audit or supervision.

### AI Act control cycle

**From risk to continuous evidence throughout the entire life cycle**



**Interpretation:** the AI Act turns AI governance into an operational capability. It is not enough merely to declare compliance; it must be demonstrated with continuous evidence.

#### Keys

1. Classify before building
2. Control before deploying
3. Demonstrate throughout the service life

**Complying with the AI Act requires operational evidence, not just corporate policies.**



# 02



## Technological evolution, models and infrastructure



Seizing the opportunity

The new  
European  
frontier of B2B AI

Strategic Intelligence Report



# From mass AI to industrial AI

The evolution of AI in regulated sectors shifts the focus from maximum power towards economic efficiency, operational control, and task-specific specialisation.

- » **The criterion is no longer brute force.** The adoption of foundation models is no longer decided by choosing the most capable model, but by identifying the right-sized model for each process, with the appropriate infrastructure, the required latency and the level of control that the data requires.
- » **LLMs are no longer the default option.** They remain relevant for complex, open-ended, or high-unit-value tasks, but no longer justify their indiscriminate use in regulated and repetitive processes, where the additional cost does not always translate into a proportionate operational improvement.
- » **SLMs come to occupy an industrial layer.** Their lower inference cost, local deployment capability and adaptation to specific tasks make them a natural choice for high-volume workflows, such as document classification, internal RAG, preliminary validations or middle-office support.
- » **Financial analysis reorganises the architecture.** As inference scales, API-based consumption turns AI into variable OpEx. In contrast, controlled deployment reduces vendor lock-in, improves budget predictability and enables infrastructure to be amortised across recurring workloads. The European ecosystem, with initiatives such as ALIA and Salamandra, provides multilingual capabilities, transparency and greater technological sovereignty.
- » **The next phase will be multimodal.** Adoption will no longer be limited to text. Voice, vision and multimodal models will enable the automation of document processes, call analysis, customer service and the processing of scanned documents with criteria of traceability, security and operational control.

# The end of “One Model Fits All”

One model is no longer enough for all tasks

**Regulated organisations need a multi-layer architecture that assigns capacity, cost and control according to the criticality of the process.**

The **One Model Fits All** approach loses its meaning when AI moves from pilot to operation. A giant model can solve many tasks, but not always with the cost, latency or level of control required by a recurring banking process.

The alternative is not to replace all **LLMs**, but to reserve them for uses where their additional capability justifies the extra cost, such as complex reasoning, open-ended tasks or high-value decisions. For narrowly scoped processes, **SLMs** offer a more efficient

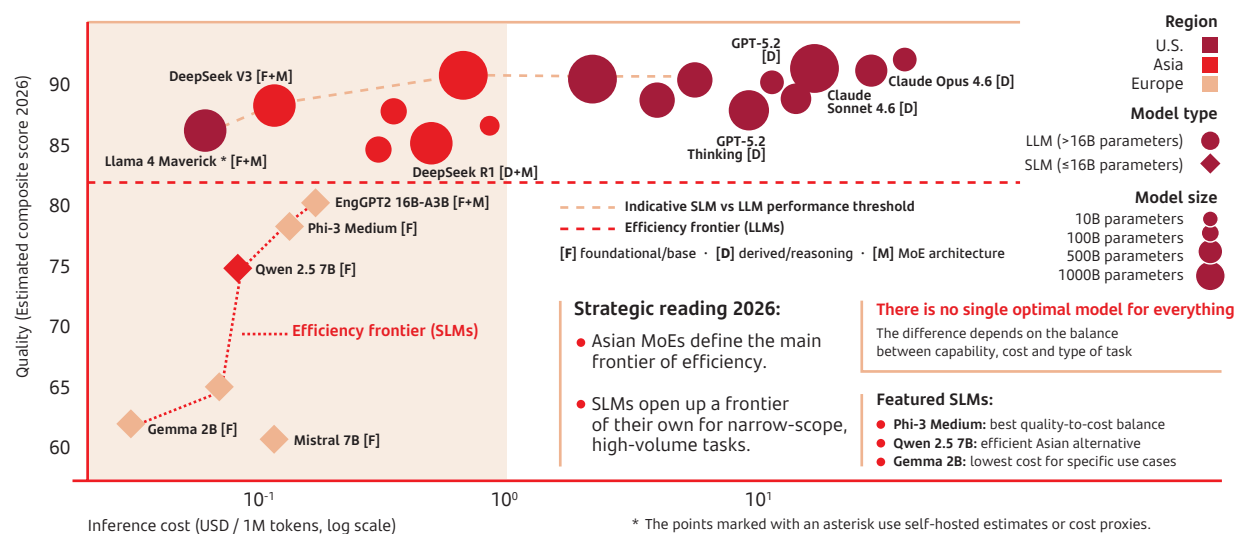
relationship between useful performance and operating cost.

With this approach, the architecture is segmented by the criticality, volume, and sensitivity of the data. The result is reduced **vendor lock-in**, improved process reconstruction, and more disciplined spending allocation, where each model occupies the layer that best fits the process.

**Enterprise architecture must adapt the model to the process, not the process to the model.**

## Efficiency frontier in foundation and derived models (2026)

Strategic assessment: LLMs, MoE architectures and specialised SLMs



# Small Language Models / SLMs

## SLMs for high-volume banking processes

**SLMs turn repetitive tasks into more economical, faster inference that can be deployed closer to the data.**

An **SLM** should not be interpreted as a smaller version of an LLM, but rather as an architecture designed for narrow-scope, repeatable and measurable tasks. Its value becomes apparent when the required capability is well defined and sufficient performance matters more than general-purpose breadth.

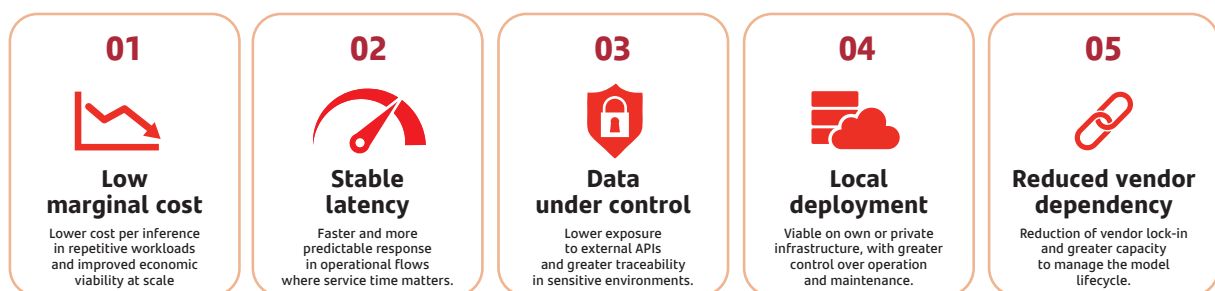
In banking, this approach is well suited to high-volume processes such as document classification, field extraction, internal RAG, preliminary validations, operational translation, or middle-office support. In these cases, the difference in quality compared with a larger model may not justify the additional inference cost.

The operational advantage is reinforced when the model can be deployed in self-hosted or private environments, close to the data or on private infrastructure. This reduces latency, stabilises costs, improves traceability and limits the exposure of sensitive information compared with architectures dependent on external APIs.

Local or self-hosted deployment reduces data exposure, stabilises latency, and converts part of the recurring expenditure into amortisable capacity. The result is an AI that is less visible, but more suitable for industrialisation.

### Why an SLM scales better in banking

#### Five operational keys for large-scale, recurring and data-sensitive processes



**READING:** the SLM becomes more relevant when the task is well defined, the volume is high, and operational control shapes the business case.

**A cost-effective SLM can outperform an oversized LLM in large-scale production.**

# European and multilingual ecosystem

## Operational linguistic sovereignty

**ALIA, Salamandra and European models make it possible to cover co-official languages with greater traceability, openness and regulatory control.**

Europe is beginning to build its own layer of open, multilingual models to reduce technological dependence and adapt AI to regulated contexts. In Spain, **ALIA** provides linguistic and multimodal resources in Spanish and co-official languages, with a focus on transparency, control and business deployment.

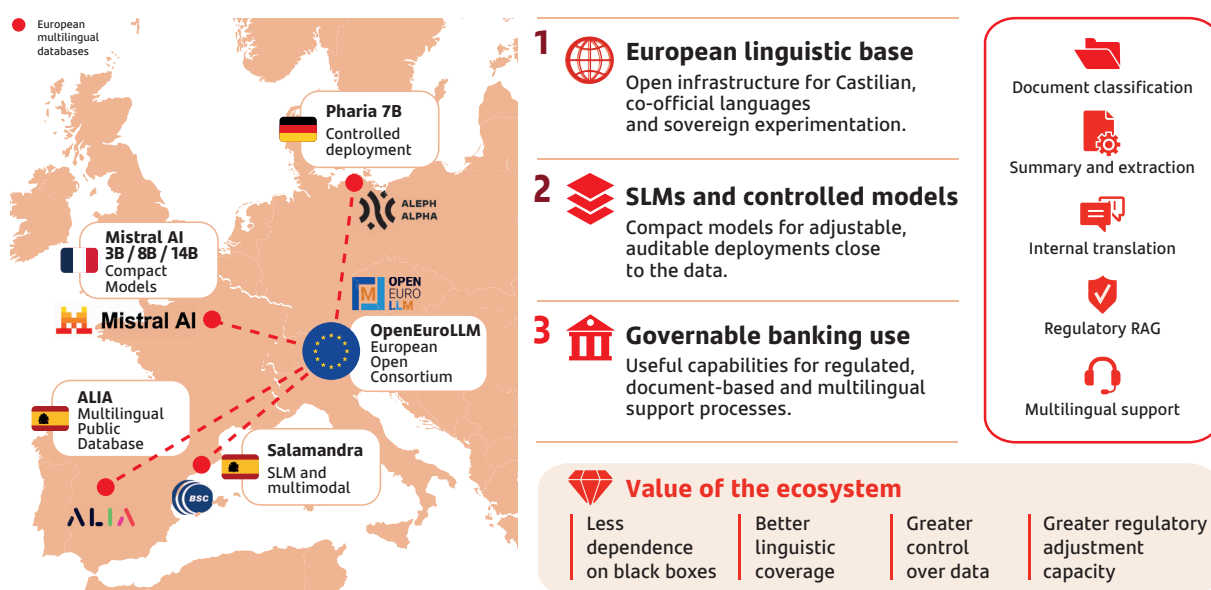
Within this framework, **Salamandra\*** functions as a family of compact models geared towards practical uses: classification,

summarisation, document extraction, translation and multimodal pilots. Its value lies not in competing with global LLMs, but in enabling adjustable, traceable capabilities that are close to the data.

In the banking context, these initiatives provide a useful basis for **exploring sovereign AI** in documentary, regulatory and multilingual customer service processes, reducing opaque dependencies and improving governance.

## European model ecosystem for banking

### Linguistic foundation, controllable models and banking applications



\* Salamandra is a family of open AI models developed within the BSC environment, focused on Iberian and co-official languages. It includes language models and multimodal capabilities, of particular interest for uses where greater transparency, local adaptation and data control are required.

**Useful sovereignty is the kind that reduces operational risk, not just technological dependence.**

# B2B Build vs Buy financial analysis

## Buy for agility, build for scale

**Build doesn't mean training from scratch; it means orchestrating open models under your own control when the volume justifies it.**

The financial debate must move beyond the false dichotomy between creating models from scratch and consuming APIs. In banking, **Build** means integrating, adapting and operating open or compact models on in-house or private infrastructure, with a focus on control, amortisation and cost stability.

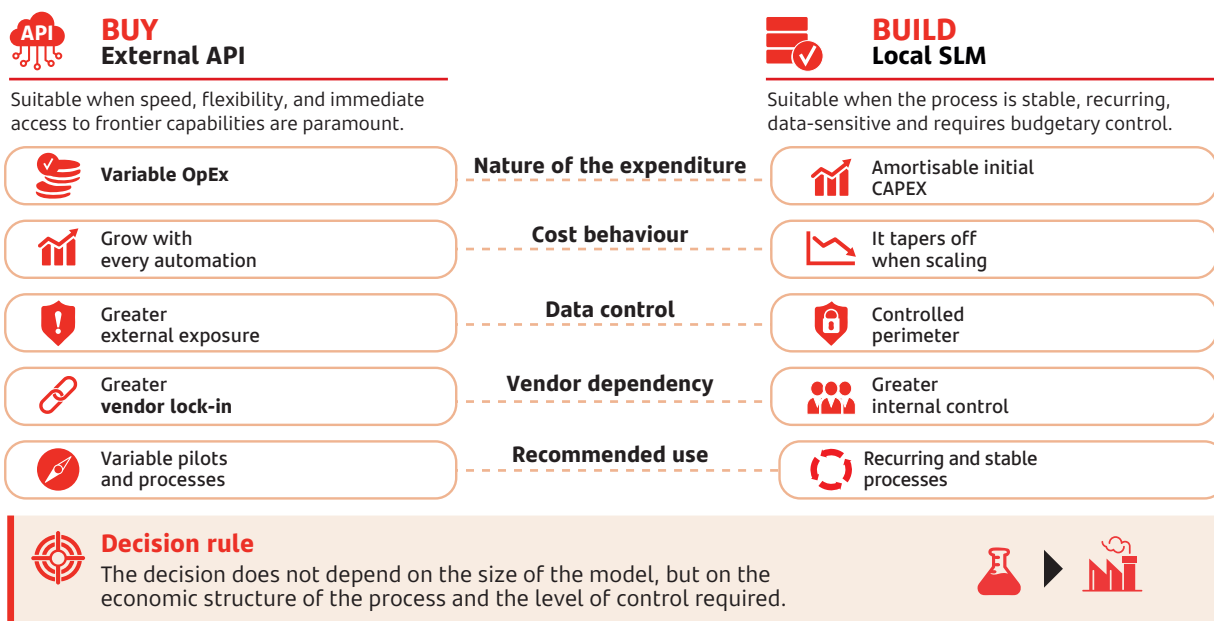
**Buy** remains useful when the organisation needs speed, elasticity or immediate access to frontier models. It works well in pilots, proof-of-concept trials and processes that are still

limited. Its limit becomes apparent when those uses cease to be occasional and each new automation increases recurring expenditure.

The decision criterion must be based on the economic performance of the process. If the volume is stable, the data is sensitive and there is internal capacity to operate the model, in-house deployment carries more weight. The most robust strategy will be hybrid, but with a clear rule on what is purchased and what is brought in-house.

### Build vs. Buy in banking

**How to decide between an external API and self-hosted deployment based on cost, control and scale**



**Buying accelerates experimentation; building provides control at scale.**

# The Total Cost of Ownership / TCO

## The token hides the cost

**TCO<sup>\*</sup> reveals when the API stops being flexible and becomes variable OpEx that is difficult to contain.**

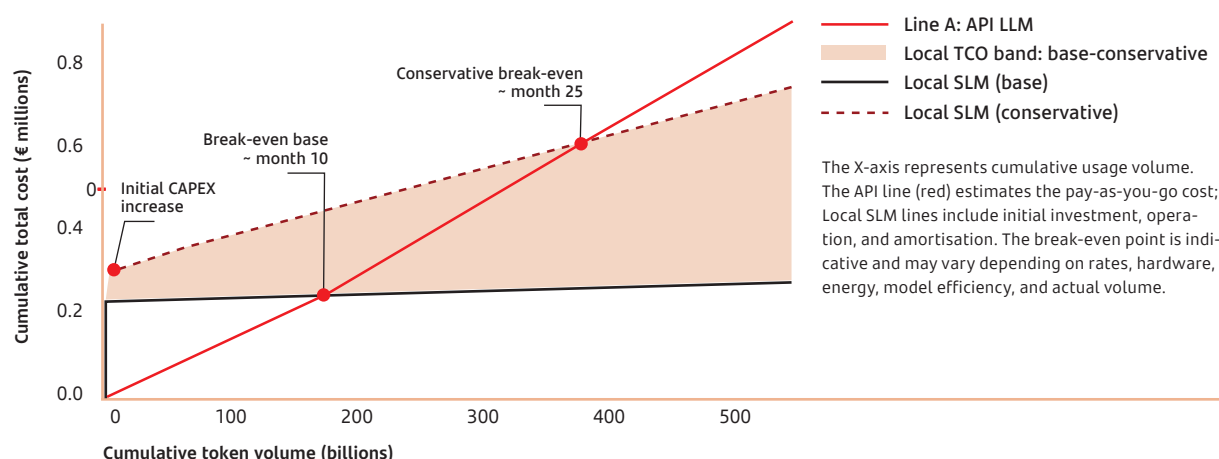
The price per token is an incomplete metric for deciding on an AI architecture in banking. It may be useful for comparing providers, but it does not explain the real cost of sustaining a process in production, with millions of inferences, security controls, monitoring and operational continuity.

With API consumption, each increase in usage becomes **variable OpEx**. The entity gains initial speed, but the accumulated expenditure remains linked to the volume processed, to the supplier's rates and to external

pricing and contractual terms. As the process scales, the initial flexibility can turn into recurring budgetary pressure.

In-house deployment requires investment, integration and operation, but it changes the cost curve. By bringing infrastructure in-house and operating **SLM** in controlled environments, expenditure becomes more predictable and can be amortised. TCO makes it possible to decide when that investment pays off in terms of volume, data sensitivity, and governance requirements.

### 36-month TCO: API vs. local SLM (baseline and conservative)



<sup>\*</sup> The TCO comparison should be interpreted as an indicative simulation, not as a fixed result. The breakeven point between API and self-hosting can shift rapidly depending on inference pricing, hardware costs, energy costs, actual usage volume, model efficiency, and operational costs. Therefore, the analysis should be reviewed periodically and maintain architectural flexibility to combine APIs, private cloud and in-house deployments as each use case evolves.

**The relevant cost is not the token, but the slope of cumulative expenditure.**

# Task specialisation

## Small models, critical tasks

**A 7B or 8B parameter model may be sufficient when the task is well-defined, documented, and supervised.**

For well-defined banking tasks, a **7B- or 8B-parameter SLM** may be sufficient if it operates on reliable sources, clear instructions, and defined validation criteria.

Performance depends not only on the model chosen, but also on how it is integrated into the business process. A model may have strong core capabilities, but its value increases when it works with reliable sources, clear instructions, and well-defined decision pathways. In banking, where precision,

monitoring and compliance are critical; the quality of the operational solution usually carries more weight than the standalone capability of the language engine.

This logic is particularly well suited to the middle office, contracts, files, internal regulations, or document translation. They are repetitive, cost-sensitive processes with sufficient structure to be industrialised. There, the gain comes from assigning the right model, not from oversizing the architecture.

### Which model is suitable for each banking task

#### Task-model map for specialised processes

|  Banking task                |  Sufficient model |  Why it fits |  Operational control |
|-----------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------|
|  Internal RAG                | SLM 7B-8B                                                                                            | Enquiry about controlled sources                                                                | High                                                                                                      |
|  Documentary translation     | SLM 7B-14B                                                                                           | Limited and repeatable task                                                                     | Medium-high                                                                                               |
|  Classification of contracts | SLM 7B-16B                                                                                           | Extraction and structured classification                                                        | High                                                                                                      |
|  Middle-office documentary   | SLM 8B-14B                                                                                           | High volume and low variability                                                                 | High                                                                                                      |
|  Human validation            | SLM + workflow                                                                                       | Monitoring at critical points                                                                   | Critical                                                                                                  |



#### Assignment principle

**Limited task + controlled source + defined validation = sufficient SLM**

In repetitive banking processes, useful performance depends more on operational design than on the model's maximum size.

*Indicative conceptual map based on documentary and middle-office tasks described in the report.*

**Specialisation turns limited size into sufficient performance.**



# Multimodal AI and speech models

## Voice and document under control

**Specific multimodality automates calls, scanned documents and forms without always relying on giant models.**

The next layer of efficiency will not be solely textual. In banking, a significant part of operational information arrives in unstructured formats, such as calls, scanned PDFs, forms, supporting documents or documentary images. **Multimodal AI** handles those materials within traceable and supervised workflows.

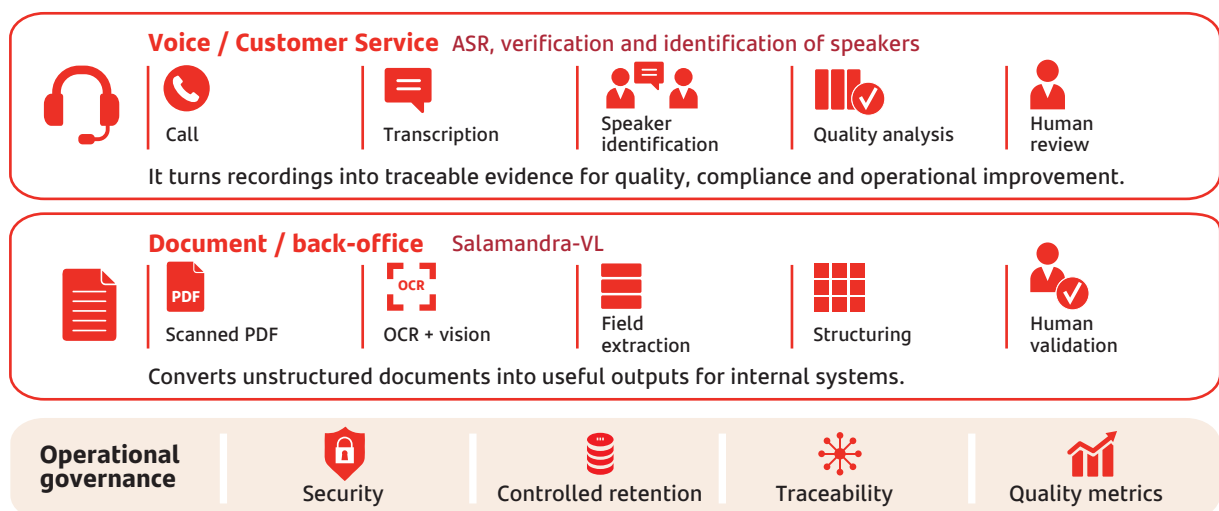
Voice and vision models are valuable when applied to specific tasks. The speaker identification and diarisation in calls, transcription, analysis of scanned documents or extraction of fields do not always require the largest general-purpose model.

They can be resolved with specialised models, integrated into controlled processes.

The priority is not to add more channels, but to manage them. In customer service, document back office or compliance, multimodality must operate with security, controlled retention, human review and quality metrics. There, models such as **Salamandra-VL** or European voice solutions such as **Whisper Large V3** deployed in on-premises environments can reduce latency and **OpEx** without losing operational control.

### Multimodal flows under control

#### Voice and document in supervised banking processes



**Operating principle:** Multimodality adds value when it is integrated into specific, measurable, and controlled workflows.

It is not a question of adding more channels, but of governing voice, document, and review within the banking environment.

# Control architectures and the agentic frontier

From sovereign RAG to supervised reasoning.

The operational value of banking AI depends on optimising deployment, protecting corporate memory and governing advanced reasoning.

- » **Efficiency is no longer determined solely by the size of the model.** Techniques such as MoE, distillation, LoRA and compression make it possible to reduce compute, latency and infrastructure consumption, bringing advanced capabilities closer to banking environments where cost, scalability and governance are determining factors.
- » **Architecture becomes the main differentiating factor.** Two organisations may use similar models, yet achieve very different results depending on how they manage information retrieval, security, traceability, integration with internal systems, and operational control.
- » **Sovereign RAG separates the language engine from the bank's document memory.** This decision prevents corporate knowledge from becoming embedded in the model and keeps policies, procedures, and critical documentation under the entity's direct management.
- » **Data sovereignty demands more than private repositories.** Zero Egress, secure enclaves and confidential computing mechanisms ensure that sensitive information remains within authorised perimeters throughout the inference and usage cycle.
- » **The leap to system 2 introduces a new generation of capabilities focused on reasoning.** Models are no longer limited to predicting responses and are beginning to verify steps, cross-check information, and justify processes before issuing a recommendation.
- » **Agentic AI expands this approach by allowing systems to interact with tools, applications and workflows.** Its adoption in banking will depend less on technological autonomy and more on the ability to impose limits, record decisions and maintain human oversight at critical points.

# Optimisation architectures

## Lighter and more useful inference

**Optimisation techniques reduce cost, latency and infrastructure consumption without turning every deployment into a heavy engineering exercise.**

Optimisation seeks to decouple functional capacity and operating costs. Not all tasks require activating the entirety of a model or maintaining oversized infrastructure. MoE architectures activate only the components required for each query, while distillation and efficient fine-tuning transfer specific capabilities to more compact, specialised models.

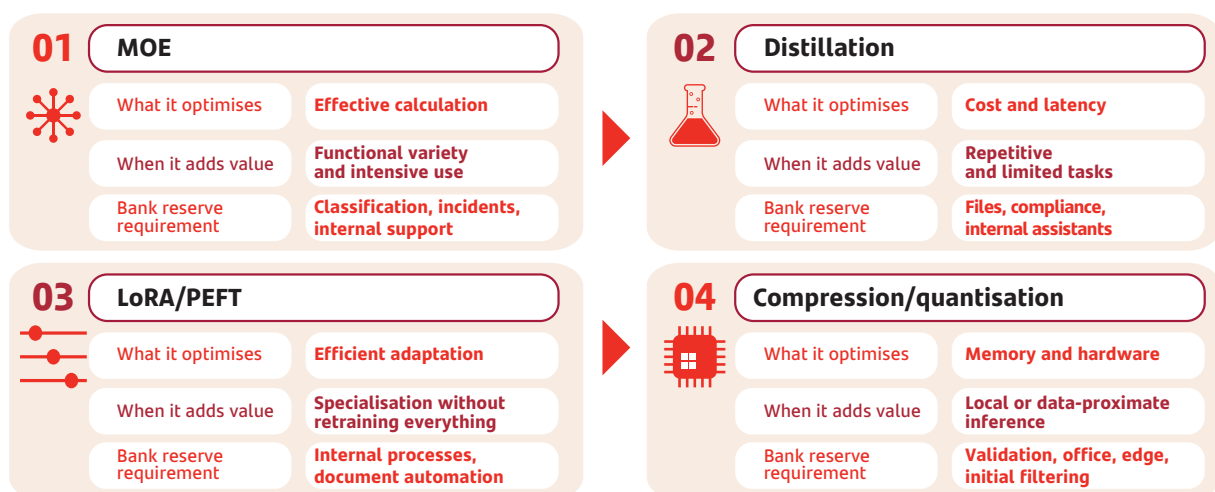
In banking, this approach brings AI closer to the data and deploys it on standard servers, internal nodes or infrastructure-constrained environments.

Techniques such as **LoRA**, compression and quantisation reduce memory, energy consumption and latency without compromising performance when the process is well defined and has controlled sources.

The key decision is not to choose an isolated technique, but rather which limitation determines the use case. If the challenge is cost, response speed or deployment capacity, the architecture should prioritise optimising system efficiency before increasing the size of the model.

### Architectures for deployable inference

#### Techniques to reduce the cost, latency, and material friction of deployment



**Interpretation:** Each technique solves a different bottleneck: cost, adaptation, latency, or local deployment.

**Optimising does not reduce ambition;  
it reduces the material friction of deployment.**

# Sovereign RAG\*

## Memory outside the model

**Sovereign RAG connects AI and corporate knowledge without incorporating sensitive documentation into the model's weights.**

Sovereign RAG introduces a deliberate separation between reasoning capability and corporate knowledge. Instead of incorporating internal documentation through continuous retraining, the model consults document repositories governed by the entity in real time. The architecture reduces dependence on the model's weights and ensures that the information used is always the current, authorised version.

The advantage is not only technical, but also operational and regulatory. By keeping knowledge outside the model, it becomes possible to update policies, withdraw obsolete documents, segment access, and reconstruct precisely which sources supported each response. Corporate memory remains under document control, with governance criteria equivalent to those applied to any other critical asset of the organisation.

For a regulated organisation, this approach makes it possible to use regulations, contracts, internal procedures or risk criteria without exposing them to training processes or losing traceability over their use. The quality of the response no longer depends exclusively on the model and instead comes to rely on the combination of reasoning, document retrieval and control of the sources consulted.

### Corporate memory outside the model

Risk of direct training versus sovereign RAG architecture



#### Risk of incorporation into the model



1 Sensitive documentation



2 Fine-tuning / training



3 Model with implicit memory



4 Difficult to reconstruct response

- More costly update
- Less clarity regarding the effective documentary basis
- More difficult subsequent reconstruction



#### Sovereign RAG



1 User query



2 Semantic search



3 Private document repository



4 Retrieved passages



5 AI models



6 Traceable response

- Greater control over data
- More agile document updates
- Better traceability of the response



**Interpretation:** Separating reasoning and storage does not eliminate risk, but it improves control, updating, and document traceability.

\* Sovereign RAG for governed corporate knowledge.

**Sensitive knowledge should be retrieved, not embedded in the model.**

# Data containment and protection architectures

The data remains within the perimeter

**Zero Egress** combines isolated networking, infrastructure, and computing to contain the data throughout the inference cycle.

**Sovereign RAG** only reaches its full potential when it is supported by an architecture capable of controlling the complete flow of information. **Zero Egress**, a data and communications isolation architecture, responds to this need by establishing a simple principle: sensitive data must not leave the authorised perimeter during retrieval, contextual enrichment, inference, or response generation processes.

Secure enclaves and confidential computing reinforce this protection during the active use of the data. The architecture separates document retrieval, storage and inference to reduce exposure surfaces, limit unnecessary

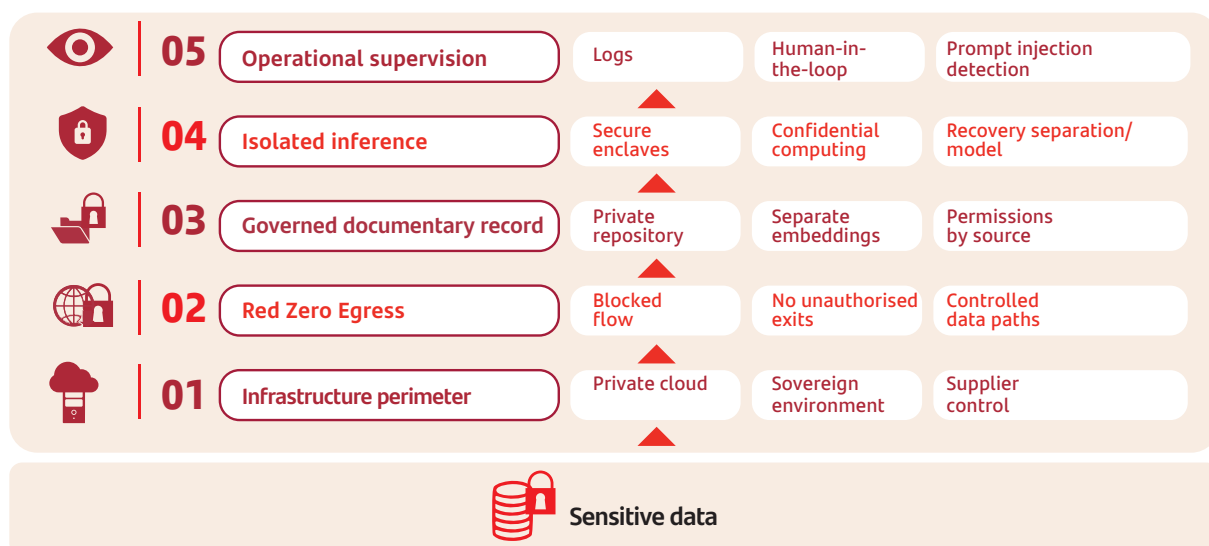
movement of information and facilitate granular control of permissions, access paths and operational logs.

In a banking environment, this layer is not merely a security measure. It is a management mechanism that shows where the data has been, who has accessed it, and under what conditions it has been processed. When technology providers, hybrid infrastructures or specialised models are involved, the ability to maintain traceability, segregation and technical oversight becomes an operational and regulatory requirement.

**Data sovereignty also requires controlling its technical journey.**

## Sensitive data containment layers

### Protection architecture for operating AI on confidential information



**Interpretation:** Data sovereignty does not depend on an isolated component, but on an architecture of networking, memory, inference and supervision.

# The leap to System 2\*

## Reasoning before responding

**System 2 adds inference, verification and contrastive computation to move from a probable answer to an explainable output.**

System 2 is not intended to replace current models, but rather to add a more deliberative layer of reasoning on top of probabilistic generation. Rather than providing an immediate response based on statistical patterns, these systems can devote more computing capacity during inference to breaking down problems, evaluating alternatives, and verifying each step before reaching a conclusion.

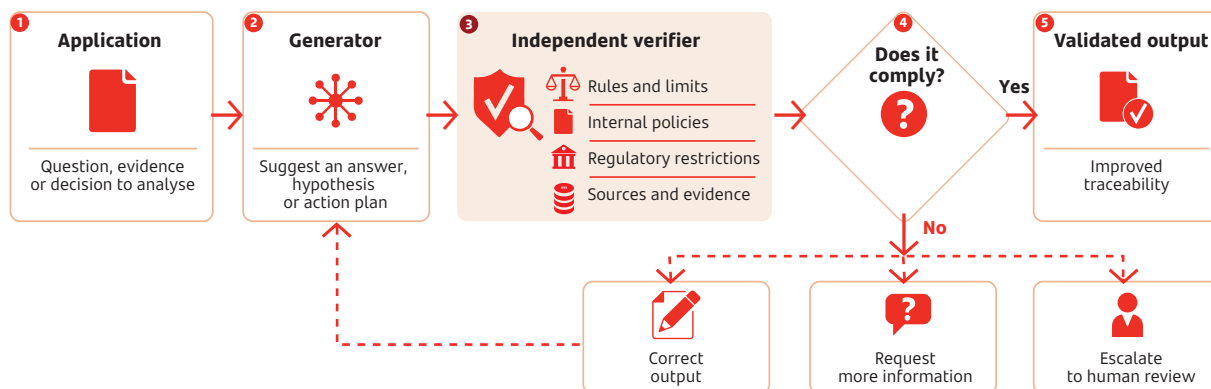
This capability is particularly useful in environments where there are explicit rules, procedures, and validation criteria. In banking, it can add value in regulatory compliance, review of internal policies, document analysis,

case management or support for risk processes, where it is not enough merely to respond, but it is necessary to justify how a particular recommendation has been reached.

Its main contribution is not to achieve absolute certainty, but to improve the quality of the decision-making process. By incorporating mechanisms for cross-checking, documentary evidence and intermediate validation, **System 2** reduces ungrounded responses, facilitates traceability and improves the capacity for human oversight in contexts with stringent regulatory requirements.

### From probable answer to verifiable output

#### Generator-verifier architecture for banking processes with rules, sources and oversight



**Interpretation:** System 2 does not eliminate risk, but it turns the response into a more traceable, reviewable and governable process.

\* The concept of System 2 is inspired by the distinction popularised by Daniel Kahneman between fast thinking and deliberative thinking. In AI, it is used analogously to describe architectures that incorporate reasoning, verification and cross-checking steps before generating a final output.

**System 2 turns the response into a verifiable process.**

# Agentic AI and complex automation

## From responding to acting

**Agentic AI turns instructions into operational workflows, provided that tools, permissions and oversight are clearly defined.**

**Agentic AI** represents a step beyond traditional conversational assistants. While these generate responses from a request, an agent can plan actions, use tools and coordinate different steps to achieve a specific goal. Its contribution consists not only in responding better, but in transforming an instruction into an executable and verifiable operational process.

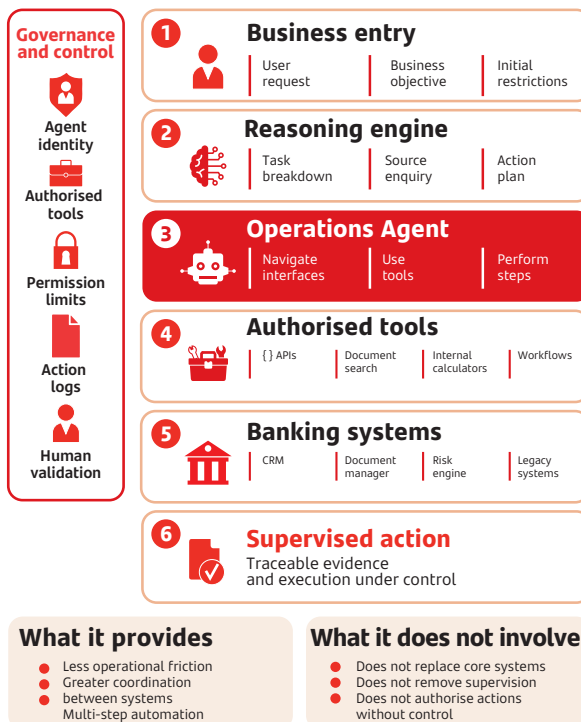
In banking, this capability opens the door to handling processes that today require multiple systems, queries and validations. Activities such as client onboarding, mortgage management, certain back-office tasks, or support for risk and compliance functions can benefit from agents capable of consulting authoritative sources, gathering information, and preparing evidence for review.

The logic is extended with multi-agent systems, where different specialised functions collaborate to solve a complex task. The division of labour can improve efficiency and scalability, but it requires robust coordination and traceability mechanisms to reconstruct each step carried out.

The adoption of Agentic AI should not be understood as a pursuit of total autonomy. Its value lies in increasing operational capacity within defined limits, combining automation, control and points of human intervention when the criticality of the process so requires.

### From instruction to governed operational flow

How agentic AI connects to tools, internal systems and banking controls



**Useful autonomy is that which operates within limits, not that which operates without control.**



**Autonomy only adds value when every tool, permission, and action is clearly defined, recorded, and supervised.**

# AI agent management

## Agents with an operational perimeter

**For an agent to operate safely within an organisation, it is necessary to clearly define who is acting, which tools they may use, what their limits are, and when a person must intervene to supervise or validate the process.**

When an **AI agent** can interact with systems, applications or processes, the risk no longer lies only in the quality of the response, but in **the action it carries out**. A recommendation can be reviewed, whereas an improperly authorised transaction can affect clients, internal controls or regulatory requirements.

Therefore, agentic management must focus on the **operational runtime**. Each agent needs a verifiable identity, a closed catalogue of tools, minimum permissions by function, and explicit limits on what it can consult, modify, or escalate. Autonomy is only acceptable if it operates within a **previously defined perimeter**.

**Traceability** is the second requirement. The bank must be able to **reconstruct what the agent did**, which source it consulted, which tool it used, which decision it proposed, and at what point a person intervened. Without complete logs and human validation in sensitive processes, the agent ceases to be governable.

In banking, trust is not based on assuming that the agent will act correctly, but on being able **to verify every step**. Efficiency comes from automating tasks, while control depends on restricting permissions and leaving an operational trail.

### Minimum controls for the agentic runtime

Operational governance checklist for AI agents in regulated banking

1



#### Agent identity

Who is acting, in what role and in what context.

2



#### Authorised tools

Closed catalogue of APIs, systems and permitted interfaces.

3



#### Permission limits

Minimum necessary access and actions restricted by criticality.

4



#### Action logs

Complete log of steps, queries, tools and results.

5



#### Human validation

Mandatory approval for sensitive or irreversible operations.

6



#### Interruption and reversal

Ability to stop, block or reverse non-compliant actions.



#### Governance principle

Every agent must operate with defined identity, permissions and tools before executing an action.



**An agent is only governable if identity, tools, permissions and actions are bounded by design.**

**The agent is governable only if identity, permissions, and actions are recorded.**



# Deep Tech radar – Strategic map of emerging capabilities

## Seek better solutions

**Evolutionary AI introduces cycles of generation, evaluation and selection to optimise solutions under explicit constraints.**

The next frontier is not only about responding better, but also about exploring solutions that had not initially been considered. Unlike conventional generative models, **evolutionary approaches generate multiple alternatives**, evaluate them using metrics or simulations, and retain the best-performing ones in order to continue refining them. More than predicting, they seek, test and optimise.

Projects such as AlphaEvolve, FunSearch and GNoME show how this approach makes it possible to explore very large solution spaces and identify options that might go unnoticed by conventional methods. Its value lies not only in generating alternatives, but also in systematically learning which work best and under what conditions.

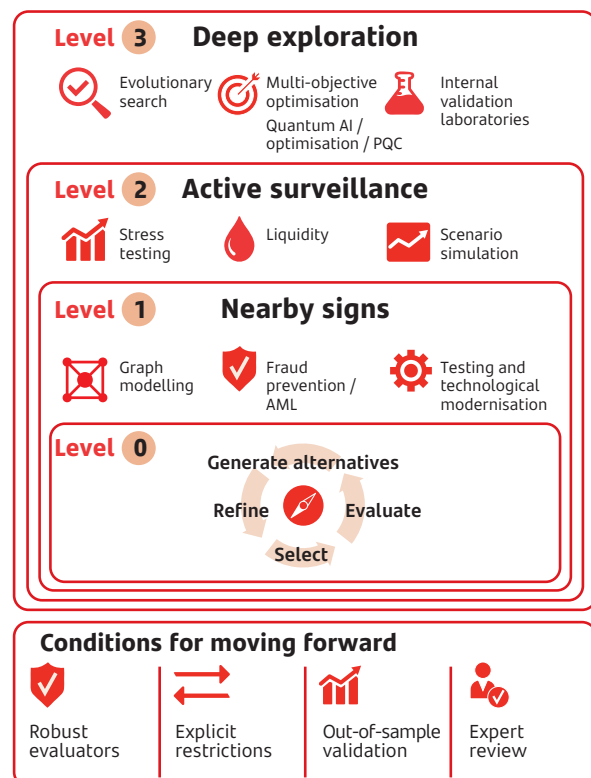
In regulated companies, these capabilities may be relevant but require caution. Risk modelling, portfolio optimisation, fraud detection, or stress-testing exercises could benefit from advanced exploration and simulation mechanisms, provided that they maintain transparent, auditable evaluation criteria aligned with existing control frameworks.

Potential emerges when there is reliable data, clearly defined objectives and robust mechanisms for evaluating results. In that context, evolutionary AI expands teams' analytical capacity, exploring scenarios and strategies that are difficult to review manually, without replacing responsibility or expert judgement.

Within this technical agenda, Quantum AI, optimisation and PQC must be maintained as active monitoring areas, with a focus on simulation, financial optimisation, cryptographic resilience and early PQC preparedness.

### Deep Tech Radar for Evolutionary AI

From prediction to the controlled search for solutions under constraints Level 3



It is not a fixed roadmap, but rather a monitoring map for problems with sufficient data, constraints, and metrics.



**Interpretation:** The value lies not in delegating the final decision, but in expanding the bank's ability to explore better options under control.

**Evolutionary AI does not automate decision-making; it expands the search for better solutions under control.**

# Sovereign infrastructure and sustainability

The physical underpinning of sovereignty and green AI

- » **AI is no longer just a matter of models.** Its deployment depends on computing capacity, data centres, energy, cooling, storage and data governance.
- » **Europe is advancing with EuroHPC, AI Factories and the future GigaFactories** to reduce technological dependence and make supercomputing an accessible capability for industry, science, public administrations and strategic sectors.
- » **Spain is gaining prominence as an operational hub** with the RES, **MareNostrum 5**, the BSC AI Factory and CESGA's new AI Factory, combining European scale, sector specialisation and territorial reach.
- » **Banking can adopt a selective orchestration strategy**, maintaining the traditional cloud for day-to-day operations and relying on sovereign supercomputing infrastructures for critical workloads, advanced simulations, and fine-tuning models.
- » **Sustainability is becoming part of the technology decision.** The **TCO** of AI must incorporate electricity, carbon, water, large-scale inference, and Scope 3 reporting, making energy efficiency and climate traceability key competitive factors.

# Factories, gigafactories and data labs

## European computing at the service of industry

**AI Factories turn supercomputing into industrial access, while data Labs prepare sensitive data under control.**

The AI Factories powered by **EuroHPC** represent a shift in approach compared with traditional supercomputing. They are not limited to offering computing capacity; rather, they create a **complete industrial adoption environment** that combines infrastructure, specialised support, development tools, and simplified access to accelerate AI projects in strategic sectors.

Gigafactories operate on a different scale. They are designed to train and deploy large-scale foundation models, concentrating enormous resources of power, storage and connectivity. Their role is to equip Europe with capabilities comparable to those of the major global technology hubs and to strengthen its technological autonomy.

For the financial sector, the most relevant element is the **data labs**. These environments make it possible to prepare, anonymise, validate and govern sensitive data within controlled perimeters before accessing external supercomputing resources. The result is a combination of scalability, intellectual property protection, and regulatory compliance that facilitates the secure use of advanced AI capabilities.

**Useful capability does not begin with the GPU, but with governed data.**

### From European infrastructure to useful banking capacity

**How AI Factories, data labs and data spaces turn supercomputing into secure operational capability for banking**

4



#### AI Gigafactory

Ultra-scaling for Frontier models and strategic capability.

3



#### AI Factory

Computing, technical support, testing, fine-tuning and advanced inference.

2



#### Controlled data lab

Preparation, anonymisation, validation and data governance.

1



#### European data space

Common framework for exchange and interoperability.

### Banking implications



Sensitive data under control



Protected IP



Secure industry access



**Interpretation:** Useful sovereignty depends not only on the available computing power, but also on preparing and governing the data before activating it.

# The Spanish supercomputing hub

## Spain as a sovereign node

**The RES, MareNostrum 5 and CESGA place Spain on the European map of advanced computing for AI.**

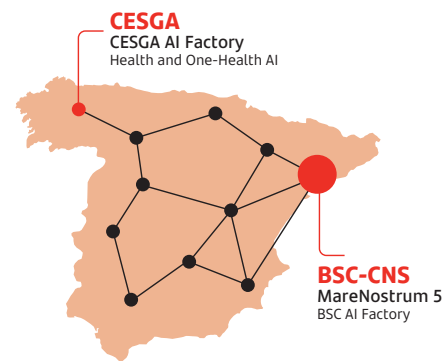
The Spanish Supercomputing Network is not an isolated facility or a single computing centre. It is a distributed infrastructure that connects fourteen specialised nodes under the coordination of the BSC. This architecture combines supercomputing capabilities, AI, advanced data management and technological experimentation within a national strategy aligned with European priorities.

**MareNostrum 5** constitutes the flagship asset of this network due to its capacity to execute highly complex workloads. Its evolution into an AI factory broadens access to advanced computing resources and facilitates the training, validation and deployment of models in controlled environments. Beyond the available power, its value lies in turning scientific infrastructure into a usable capability for businesses and institutions.

The awarding of a new AI factory to **CESGA** consolidates this position and adds specialisation in strategic areas linked to health and One-Health AI. The combination of a distributed network, international benchmark capabilities and new sector hubs positions Spain as one of the main pillars supporting European technological sovereignty.

### Spain as a sovereign node

The **RES**, **MareNostrum 5** and **CESGA** strengthen Spain's role in European supercomputing for AI



#### Connection to EuroHPC

European scale at the service of technological sovereignty



#### RES

Distributed infrastructure that connects supercomputing, AI and data capabilities at a national scale.



#### MareNostrum 5

Advanced computing asset for training, validation and inference in controlled environments.



#### CESGA AI Factory

It adds a new sectoral capability and strengthens Spanish specialisation in AI.



#### Computing sovereignty

Spain combines scale, territorial reach and operational alignment with European priorities.

**Spain is transitioning from participating in EuroHPC to operating critical AI nodes.**

# The third path for regulated sectors

## Combining cloud, on-premises infrastructure and European supercomputing

**Regulated companies do not need to replace their infrastructure, but rather to enable European supercomputing for critical and differentiating processes.**

Banking does not need to choose between relying entirely on large technology providers and reproducing internally all the computing capacity required for general AI. The most efficient approach is **to combine different infrastructures according to the value and requirements of each process**. The cloud and on-premises infrastructure remain suitable for routine operations and continuous execution, while **European supercomputing** can be reserved for those processes that require exceptional capacity or **greater strategic control**.

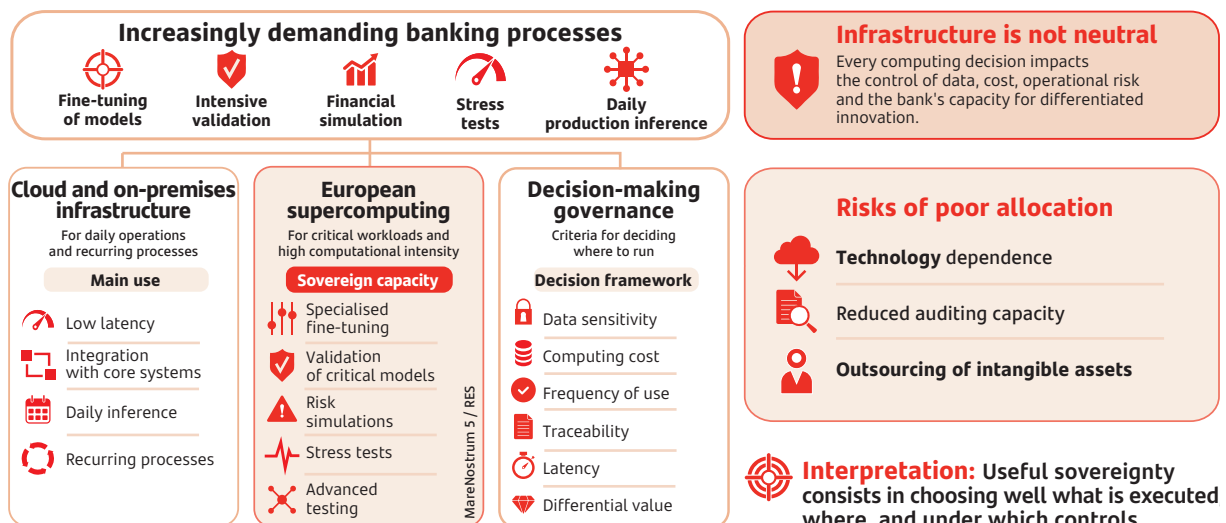
This **selective orchestration** accesses advanced resources when they genuinely provide an advantage. It is particularly useful for the fine-tuning of specialised models, intensive

validation of algorithms, complex financial simulations, regulatory stress testing or large-scale portfolio analysis. It also facilitates **experimentation with foundational models** adapted to the specific needs of the financial sector.

The decision should not be based solely on technological criteria. Each use case must be analysed taking into account the sensitivity of the data, the total computational cost, the frequency of use, traceability needs, latency requirements, and the differentiating value it generates for the organisation. The objective is to assign each process to the environment that offers the best balance between efficiency, control and technological sovereignty.

### Third route for banking computing

**The bank does not replace its infrastructure. Decide which processes are run in each layer according to criticality, cost and control.**



# The energy bottleneck

## The invisible infrastructure that supports AI

**AI scales on electricity, cooling, and water, not just on algorithms, data, and APIs.**

The expansion of AI shifts the technology debate towards a much more material reality. Behind each model are data centres (DPCs), communications networks, storage systems and cooling mechanisms that require an increasing amount of physical resources. As organisations incorporate AI capabilities into their processes, the pressure on these infrastructures increases, making **energy availability** a strategic factor.

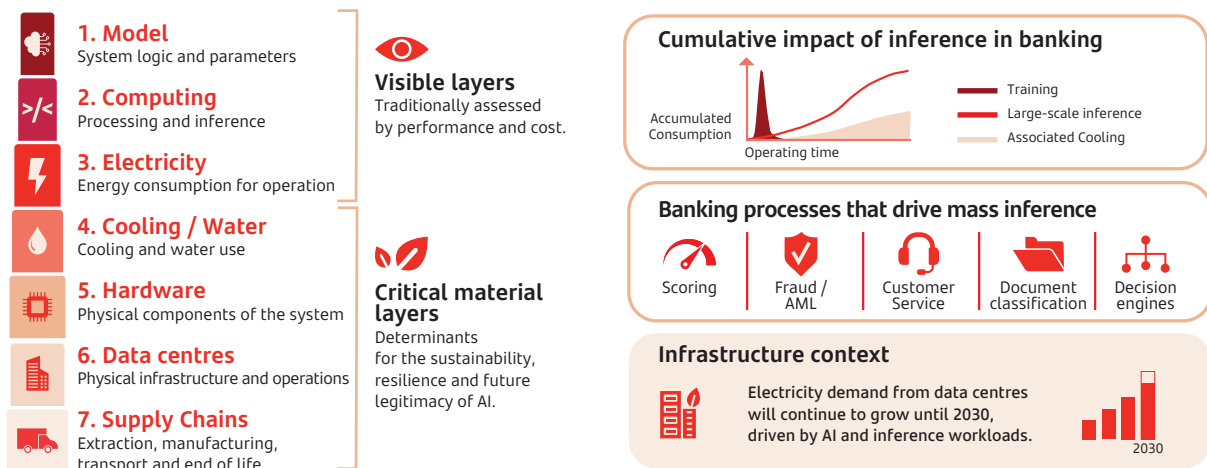
**Training** advanced models concentrates significant computing requirements during specific periods. However, the **real challenge** arises when these models go into **production**. **Inference** deployed on millions of daily interactions generates continuous demand

associated with processes such as fraud detection, risk assessment, customer service or document automation. What appears to be an individual low-impact query becomes, on a large scale, a sustained consumption of resources.

To this energy dimension are added other relevant factors such as the emissions associated with electricity consumption and the water use required for **cooling** the facilities. Therefore, the **sustainability** of AI does not depend solely on the design of the model. It is also influenced by the efficiency of the infrastructure, the location of the data centres and the quality of the energy system that powers them.

## The physical footprint of banking AI

### From the model to the continuous consumption of infrastructure, energy and water



**Interpretation:** Banking AI consumes less through large, isolated training runs and more through millions of repeated inferences each day.

# Green AI and regulatory management

## Efficient models by necessity

**Green AI connects performance, consumption and compliance, turning energy and resource efficiency into a criterion for technology management.**

Sustainable AI (**Green AI**) does not imply giving up advanced capabilities or limiting innovation. Its approach consists of **using the appropriate scale for each need** and avoiding the use of oversized models when more efficient alternatives exist. In the banking sector, a significant proportion of documentary, operational and support processes can be addressed using specialised smaller models (SLMs) that deliver comparable results with significantly lower consumption.

The approach combines different techniques to optimise resources and maintain performance. SLMs, model distillation, retrieval-augmented generation, compression, and hybrid architectures reduce energy consumption, shorten response times, and contain operating costs. Efficiency becomes part of the architectural, cost and technological governance decision.

European regulation is pushing in the same direction. Both the European AI Regulation (AI Act) and the new requirements concerning data centres are driving greater transparency regarding the material impact of models and the infrastructure that supports them. Measuring, comparing and justifying the use of resources will become increasingly important in demonstrating sustainability, regulatory compliance and long-term value creation.

### Five keys to Green AI in banking

**From technical efficiency to regulatory and financial obligation**

**01**

#### Choose the appropriate scale

Not all processes require large models.



Model

**02**

#### Prioritise specialised SLMs

Lower consumption, lower latency and better control.



Computing

**03**

#### Optimise architecture

RAG, distillation, compression and hybrid deployments.



Architecture

**04**

#### Measure material impact

Electricity, water, hardware, data centre, and life cycle.



Infrastructure

**05**

#### Document compliance

AI Act, metrics, traceability and model governance.



Governance



#### From technical efficiency to regulatory obligation

Efficiency must be measured, documented and explained to internal governance bodies and supervisors.

AI Act

CPD metrics

Energy traceability

#### Material layers considered



Model



Computing



Electricity



Cooling / water



Hardware



Data centres



Supply Chains



**Interpretation:** Green AI turns model efficiency into a decision about cost, compliance and operational legitimacy.

**The scale of AI will increasingly need to be justified by its efficiency, not what merely maximises performance.**



# The challenge of financed emissions

## Scope 3\*, the great challenge

**The largest banking climate footprint is not in offices or the bank's own technology, but in the financed economic activity.**

Bank climate reporting faces one of its greatest challenges in financed emissions. Within **Scope 3**, category 15 attributes to entities a portion of the emissions generated by borrowers, investee companies, issuers or projects that receive financing. This dimension is particularly relevant because it links financial activity with the climate impact of the real economy.

Complexity goes far beyond a methodological issue. Banks must integrate information from multiple sources, with varying levels of quality and maturity. This is compounded by incomplete data, unstructured documentation and the need to assign emission factors

consistently across thousands of exposures. In this context, AI is emerging as a tool capable of improving the identification, classification and estimation of climate information, strengthening the consistency of calculation processes.

Supervisory pressure is also increasing. It is no longer enough merely to announce decarbonisation targets or commitments. Entities must demonstrate, with verifiable evidence, how they define their calculation boundaries, the quality of the data used, and how their credit and investment portfolios are progressing in relation to climate targets.

## The challenge of financed emissions

**The bank's largest climate footprint lies in the economic activity it finances**



**“The true core of banking climate reporting lies in what is financed, not only in what is consumed.”**



### Critical data

Financed emissions can be more than **700 times** higher than the direct emissions reported by financial institutions.



### Disclosure gap

Only **15%** of the banks analysed disclosed financed Scope 3 emissions.

Incomplete data  
Heterogeneous perimeters  
Insufficient climate KPIs



**Interpretation:** Banks' climate credibility depends less on measuring offices and more on demonstrating which emissions are financed.

\* The Scope 3 concept comes from the GHG Protocol, an international emissions accounting framework that classifies indirect emissions from the value chain. In banking, this includes, among others, emissions associated with loans, investments and financed projects.



# AI for climate intelligence

## AI for quantifying financed emissions and climate risk

**AI itself can fill climate information gaps, organise ESG data and improve financial reporting.**

AI can become a key tool for addressing some of the challenges that it itself contributes to exacerbating. As climate transparency requirements increase, financial institutions must manage large volumes of information from heterogeneous sources and with varying levels of quality. In this context, AI facilitates the identification of relevant data, the classification of economic activities, the detection of information gaps, and the transformation of scattered documents into comparable and useful metrics for decision-making.

Experiences such as BIS Gaia, Symbiosis, NEMO, IBM Research and the initiatives promoted by the Bank of Spain show that these capabilities already have practical applications. AI extracts climate disclosures, structures complex corporate information, supports the application of methodologies such as PCAF, and reinforces the traceability required to respond to supervision and audit requirements.

Its contribution is particularly valuable when integrated with external data, recognised standards and expert validation processes. In this way, entities can expand their analytical capacity, improve the quality of reporting and move towards more consistent management of financed emissions and the associated climate risks.

**Climate AI adds value when it turns incomplete data into traceable decision-making.**

### Map of climate intelligence solutions

From incomplete ESG data to reporting, portfolio analysis and traceable decisions Data entry

#### 1 Data inputs

ESG documents, portfolios, transactions, counterparties and suppliers.



#### 2 AI / analytics layer

Extraction, classification, estimation, validation and gap detection.



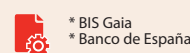
#### 3 Bank outputs

Climate reporting, Scope 3, portfolio analysis, supervision and financing.



#### 1 Document extraction and organisation

It turns reports, memoranda and climate documents into reusable data.



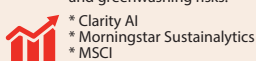
#### 2 Estimated financed emissions

Links exposures to economic activities, emission factors and methodologies.



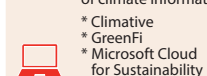
#### 3 ESG data and portfolio analysis

It assesses companies, climate transition, portfolios and greenwashing risks.



#### 4 B2B operations and reporting

It automates the collection, validation and traceability of climate information.



#### Data inputs for the bank



#### Necessary control

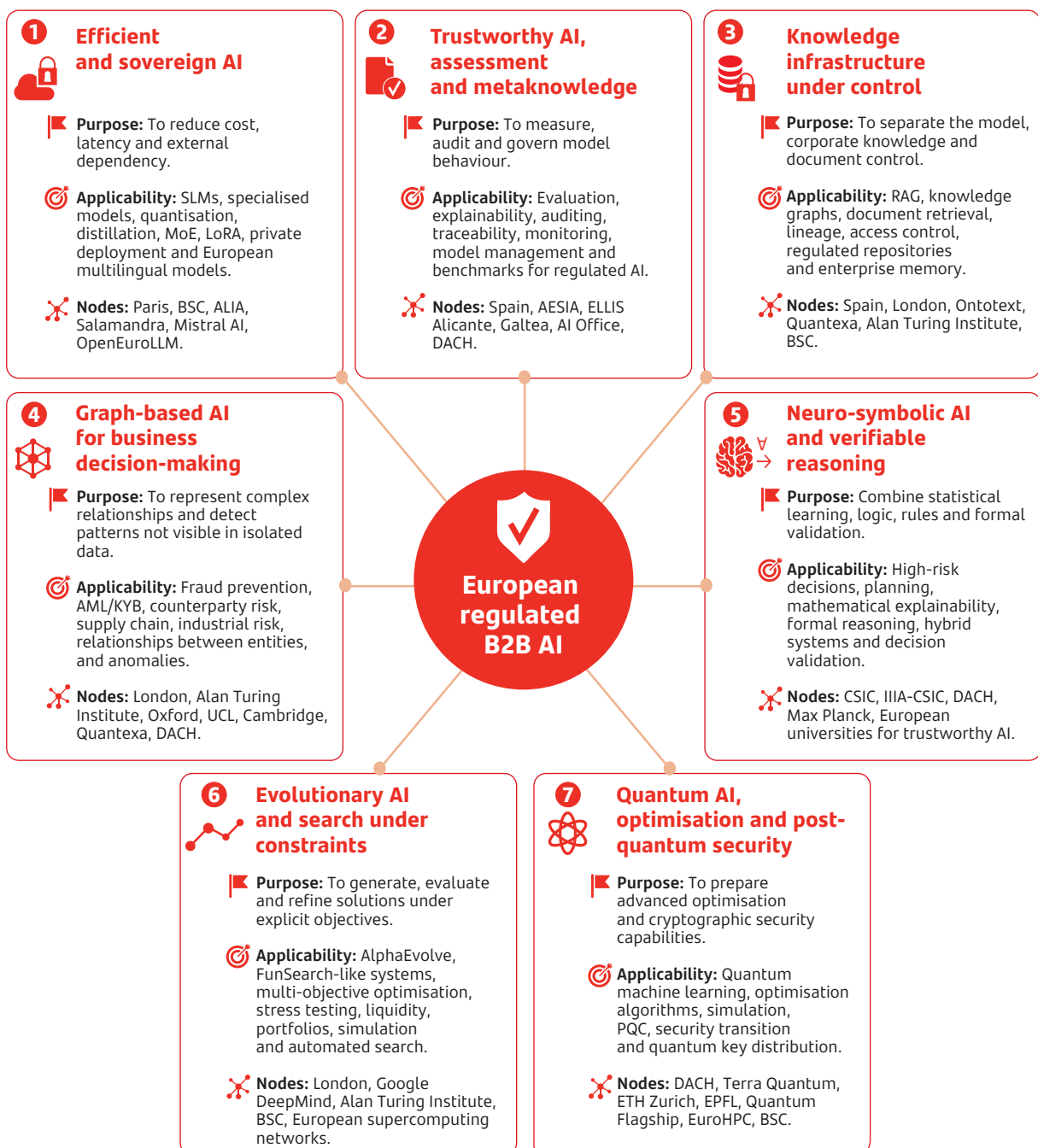
AI helps to estimate and prioritise, but it requires methodology, validation and expert review.

# Europe's priority technology agenda

## From emerging trends to foundational lines of development

New branches of AI should not only be at the heart of strategic oversight, but already form an agenda of foundational capabilities for competing in regulated B2B AI with sovereignty, traceability and industrial applicability.

They are not just trends. These are foundational lines for developing regulated, traceable capabilities that are less dependent on closed platforms.



# 03



## Regulatory framework and compliance



Seizing the opportunity

The new  
European  
frontier of B2B AI

Strategic Intelligence Report



# Compliance and regulation

The regulatory framework as a defensive advantage for scaling banking AI with control, sovereignty and supervisory responsiveness.

- » **The AI Act and DORA raise the barrier to entry for B2B suppliers.** Traceability, technical documentation and operational resilience are no longer secondary requirements, but have become key selection and procurement criteria.
- » **Compliance acts as a market quality filter.** Solutions capable of demonstrating controls, evidence and auditable processes generate greater trust and reduce regulatory uncertainty for the entity.
- » **The explainability and human oversight required by the AI Act strengthen the business adoption of AI.** Models that can justify their results are easier to validate, scale and defend in audits and before supervisors.
- » **Data sovereignty reduces legal exposure, technological dependency, and the risks associated with extraterritorial access to sensitive information.** The architecture must preserve control, confidentiality and operational continuity.
- » **Integrating AI into Model Risk Management (MRM) enables the governance of models, bias, changes, providers and oversight mechanisms under a single risk framework.** This approach protects the Bank's investment and facilitates complete traceability from design through to operation.

# Regulation as a product

## Compliance as a commercial advantage

**Compliance by design turns traceability, human oversight, and technical documentation into attributes that differentiate the solution.**

The **AI Act** changes the criteria by which a financial institution evaluates and selects AI solutions. Technical capability remains relevant, but it is no longer sufficient to justify adoption at scale. Organisations must demonstrate the origin of the data, the traceability of the processes, human oversight of decisions, and the existence of controls that make it possible to explain the system's behaviour to customers, auditors and supervisors. Transparency thus ceases to be merely a regulatory requirement and becomes an attribute of trust and differentiation.

This change reinforces the principle of **Compliance by Design**, whereby compliance is incorporated from the earliest stages of design and procurement. The decision to build in-house, buy an off-the-shelf solution, or combine both approaches should be evaluated according to the level of control it allows to be maintained throughout the entire lifecycle. The ability to audit suppliers, preserve data lineage, intervene in the models and generate verifiable evidence determines the regulatory robustness of the investment and reduces future risks of adaptation, technological dependency or non-compliance.

### Compliance as a commercial advantage

Compliance by design turns traceability, human oversight and technical documentation into product attributes.

#### Compliance by Design as a technological decision filter

| REGULATORY CRITERION                                                                                             | BUY                                                                                                      | BUILD                                                                                                                      |
|------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------|
|  Data lineage                 |  Vendor dependency  |  Greater internal control             |
|  Documentation and validation |  Inherited evidence |  Own compliance effort                |
|  Deployment time              |  Faster             |  More conditioned by validation       |
|  Technological control        |  Risk of dependency |  Greater sovereignty and adaptability |

#### Key factors for a governable decision

|                                                                                                  |                                                                                                  |                                                                                                         |                                                                                                         |                                                                                                        |                                                                                                           |
|--------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------|
|  Data lineage |  Audit rights |  Human supervision |  Reusable evidence |  Supplier control |  Complete life cycle |
|--------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------|



The Build strategy is not always preferred over Buy. Choose the most manageable option throughout the entire **lifecycle**.

**In regulated sectors, an AI that cannot be governed ceases to be a solution and becomes a risk.**

# Integrated management and MRM\*

## MRM for the safe scaling of AI

**AI in regulated environments requires a single model that integrates regulation, resilience, cyber security and technical control.**

A regulated entity should not approach the **AI Act**, **DORA** and **NIS2** as separate obligations. In practice, a single **AI solution** may trigger simultaneous requirements relating to explainability, operational resilience, cybersecurity, data protection and third-party oversight. Managing them separately increases the risk of duplication, inconsistent controls, and responses that are difficult to coordinate before supervisors and governing bodies.

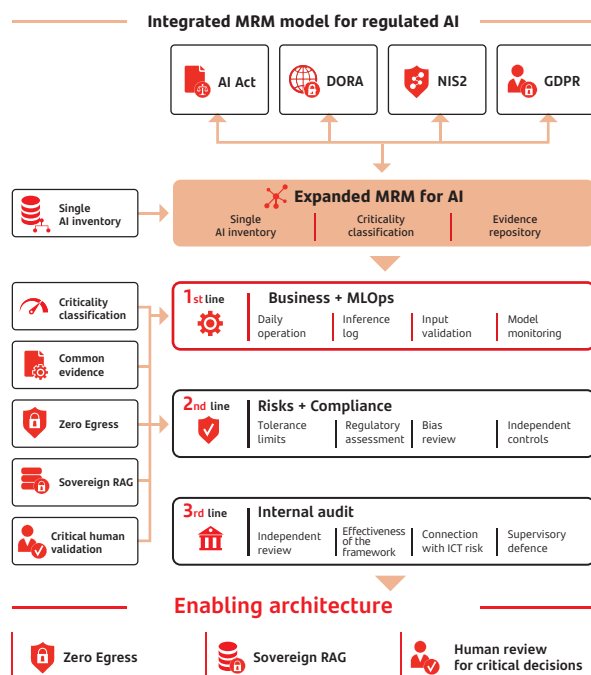
The next step is to expand the MRM framework into an integrated supervision model supported by three lines of defence. The business areas and operational teams apply continuous controls to the operation of the models. Risk and compliance functions verify that the established limits are respected and assess the regulatory and ethical impact of each use case. Internal audit provides an independent review of the effectiveness of the whole.

This approach is reinforced through architectures such as **Zero Egress**, sovereign RAG over document repositories controlled by the bank, and human validation mechanisms that strengthen traceability and responsiveness throughout the lifecycle.

### MRM to scale safe AI

Banking AI needs a single model that integrates regulation, resilience, cyber security and technical control.

**A single risk language for technology, business, compliance and audit**



**Scaling AI requires a common risk language for technology, business, compliance and audit.**

\* MRM, or Model Risk Management, is the framework that enables the identification, validation, monitoring and auditing of the models used by the Bank throughout their lifecycle.

# Current case: access to frontier models and operational sovereignty

When access to the model also becomes  
a geopolitical risk

**The Fable 5 and Mythos 5 case shows that frontier models are beginning to behave like strategic infrastructure, not just commercial software.**

In June 2026, **Anthropic** announced that the United States Government had issued an export control directive to **suspend access to its Fable 5 and Mythos 5 models** for non-US nationals, whether inside or outside the country. To ensure compliance with the order, the company stated that it had to deactivate both models for all its customers, although the rest of Anthropic's models were not affected.

A few days later, following clarification of the scope of the measure by the U.S. authorities, Anthropic restored access to both models for the affected users. Although the disruption was temporary, the episode highlighted that access to frontier AI capabilities can be influenced by regulatory and geopolitical decisions beyond the control of the organisations that depend on them.

Beyond the specific case, the episode illustrates a **growing risk for regulated**

**companies and critical sectors.** Access to frontier models may be conditional on external decisions regarding industrial policy, national security or export controls. For an entity integrating AI into critical processes, exclusive reliance on closed models subject to external jurisdictions may introduce risks to continuity, traceability, compliance, and operational sovereignty.

This type of situation reinforces the central thesis of the report. AI sovereignty does not imply isolating oneself from global models, but rather **building a hybrid, flexible and governable architecture.** This requires combining external providers, open-weight or self-hostable models in controlled environments, internal evaluation capabilities, contingency plans, and clear criteria on which processes can depend on external services and which require greater operational control.

**The critical dependency is not only in the data or the infrastructure. It is also in the continuity of access to the models that support sensitive processes.**







## Terms of use, ownership and team

The content, analyses, interpretations and conclusions contained in this report have been prepared within the framework of the strategic intelligence project on AI developed for Banco Santander. The document is intended solely for informational and strategic purposes and to support internal reflection, and should not be interpreted as a definitive recommendation for investment, procurement, technological implementation or operational adoption.

The information included in the report is based on public sources, technical documentation, expert analysis and working materials gathered during the development of the project. Although a reasonable effort has been made to verify and organise the available information, the field of AI is evolving rapidly, so the data, references, technologies, companies and conclusions included may be subject to subsequent updating or revision.

References to external companies, technologies, institutions, experts or initiatives are descriptive and analytical in nature. Their inclusion does not imply commercial endorsement, formal validation,

or a contractual relationship nor any official position on the part of Banco Santander, unless expressly stated otherwise.

The use of this report is limited to the internal and professional scope authorised by Banco Santander. Any reproduction, distribution, public communication, adaptation or total or partial reuse of its contents must have prior authorisation, except in those cases where it is expressly permitted and provided that the source is properly cited.

Ownership of this report and of the materials developed within the framework of the project belongs to Banco Santander, without prejudice to any rights that third parties may hold over the trademarks, logos, fonts, references, images, graphics or content cited. All trademarks and trade names mentioned belong to their respective owners.

This document has been prepared by the project's working team, made up of professionals from Banco Santander and the Fundación General CSIC in the tasks of analysis, content structuring, strategic intelligence and document preparation.

## Team

Banco Santander: AI Lab Team

Fundación General CSIC: Strategic Intelligence Unit (Dept. Innovation)

---

This report is available under the CC BY-NC 4.0 licence:  
<https://creativecommons.org/licenses/by-nc/4.0/?ref=chooser-v1>

---

## How to cite this report:

Banco Santander and Fundación General CSIC. (2026). *The new European frontier of B2B AI*. Banco Santander.

© Banco Santander, 2026. All rights reserved.

Corporate headquarters: CGS Ave. Cantabria s/n  
28660 Boadilla del Monte, Madrid (Spain)

