

Issue N. 32 - 2026

ARGO

New Frontiers in **Practical** Risk Management



This is a creation of **iason**.

The ideas and model frameworks described in this document are the result of the intellectual efforts and expertise of the people working at **iason**. It is forbidden to reproduce or transmit any part of this document in any form or by any means, electronic or mechanical, including photocopying and recording, for any purpose without the express written permission of a company in the **iason Group**.



Argo magazine

Year 2026 - Issue Number 32

Published in July 2026

First published in October 2013

Last published issues are available online:

<http://www.iasonltd.com>

Front Cover: **Umberto Boccioni**, *Costruzione spirale*, 1913.



ESSENTIAL SERVICES FOR
FINANCIAL INSTITUTIONS



Editors:

Antonio CASTAGNA (Managing Partner)

Luca OLIVO (Managing Director)

Executive Editor:

Lorena CORNA

Graphic Designer:

Nathalie ALARCON

Scientific Editorial Board:

Michele BONOLLO

Alessandro CAPPO

Marco CARMINATI

Giacomo COLOMBO

Dario ESPOSITO

Fabrizio GENTALAVIGNA

Sergio GRASSO

Massimo GUARNIERI

Jacopo MANCIN

Sara MARTUCCI

Andrea MAURI

Giuseppe MORISANI

Antonio MENEGON

Gianmarco MORI

Giulia PERFETTI

Andrea PRINCIPE



Milan Headquarter:

Corso Europa, 15
20122 Milan
Italy

London Headquarter:

58-60, Kensington Church Street
W8 4DB London
United Kingdom

Madrid Headquarter:

Calle de José Ortega y Gasset, 22-24
28006 Madrid
Spain

Contact Information:

info@iasonltd.eu
www.iasonltd.com

iason is a registered trademark.

Articles submission guidelines

Argo welcomes the submission of articles on topical subjects related to the risk management. The articles can be indicatively, but not exhaustively, related to models and methodologies for market, credit, liquidity risk management, valuation of derivatives, asset management, trading strategies, statistical analysis of market data and technology in the financial industry. All articles should contain references to previous literature. The primary criteria for publishing a paper are its quality and importance to the field of finance, without undue regard to its technical difficulty. Argo is a single blind refereed magazine: articles are sent with author details to the Scientific Committee for peer review. The first editorial decision is rendered at the latest within 60 days after receipt of the submission. The author(s) may be requested to revise the article. The editors decide to reject or accept the submitted article. Submissions should be sent to the technical team (info@iasonltd.eu). $\text{\LaTeX}$ or Word are the preferred format, but PDFs are accepted if submitted with $\text{\LaTeX}$ code or a Word file of the text. There is no maximum limit, but recommended length is about 4,000 words. If needed, for editing considerations, the technical team may ask the author(s) to cut the article.

Table of Contents

Editorial	p. 7
Just in Time - iason Notes	p. 9

INSURANCE RISK

Insurance Risk Management: Overview of Solvency Framework and Comparative Insights on Standard and Internal Models <i>G. Mori, M. Bazzuchi, F. Bertornati, D. Bruttomesso and A. Cavazzana</i>	About the Authors	p. 20
	Introduction	p. 23
	Regulatory Framework: From Solvency I to Solvency II	p. 23
	Solvency II: Principles and Tools	p. 32
	Solvency II: Framework Evolution	p. 40
	Deep Dive Into Standard Approach and Internal Models in Insurance	p. 46
	References	p. 56

INNOVATION

**Understanding and Governing
Agentic AI: Foundations, Risks and
Control Frameworks**
A. Menegon, V. Ciminelli and M. Ranieri

About the Authors	p. 58
Introduction	p. 60
Foundation of Agentic AI	
Approach	p. 61
Implementation as a Risk	
Decision: The Case for a	
Controllable Stack	p. 63
The Emergent Risk Landscape	
and Its Regulatory Anchoring	p. 66
From Model Risk Management	
to Agent Risk Management	p. 70
The Pillars of an ARM	p. 71
Conclusions	p. 77
References	p. 78
Sitography	p. 79

**Classifying Financial Transaction
for Invoice Matching**
M. Zanoni and E. Veksin

About the Authors	p. 81
Introduction	p. 83
The K-Nearest Neighbors	
Approach	p. 85
Classifying Transaction with LLM	p. 95
References	p. 103

DEAR READERS,

Welcome to the Summer Edition of Argo Magazine.

This issue brings together three contributions spanning Insurance Risk and Innovation, offering perspectives on the evolving relationship between risk management, regulation, and technological innovation in finance.

*We open this edition with the Insurance Risk section, featuring “**Insurance Risk Management: Overview of Solvency Framework and Comparative Insights on Standard and Internal Models**” by G. Mori et al. This contribution examines the evolution of European insurance regulation from Solvency I to the more sophisticated and risk-sensitive Solvency II framework, highlighting the growing alignment between prudential requirements and insurers’ actual risk profiles. Attention is devoted to the comparison between the Standard Formula and Internal Models for the calculation of the Solvency Capital Requirement, analyzing their respective assumptions, calibration approaches, and implications for capital efficiency and risk management. The paper also discusses recent regulatory developments, including the Solvency II Review, offering a comprehensive perspective on how solvency regulation and supervisory expectations continue to shape the resilience of the European insurance sector. The paper also discusses how the Solvency II review reinforces the need for insurers using Internal Models to benchmark their results against the Standard Formula, increasing supervisory focus on the consistency and explainability of capital outcomes.*

*In the Innovation section, we begin with “**Understanding and Governing Agentic AI: Foundations, Risks and Control Frameworks**” by V. Ciminelli and M. Ranieri, which explores one of the most significant developments in artificial intelligence: the emergence of autonomous, purposeful AI agents. As financial institutions increasingly evaluate these technologies, the authors examine the new risks associated with agentic systems and the limitations of traditional model risk management frameworks. The paper discusses the evolving regulatory landscape and introduces the Agent Risk Management (ARM) framework, a dedicated governance approach designed to oversee the entire decision-making lifecycle of autonomous AI systems. By addressing issues related to data, vendors, architecture, and human factors,*

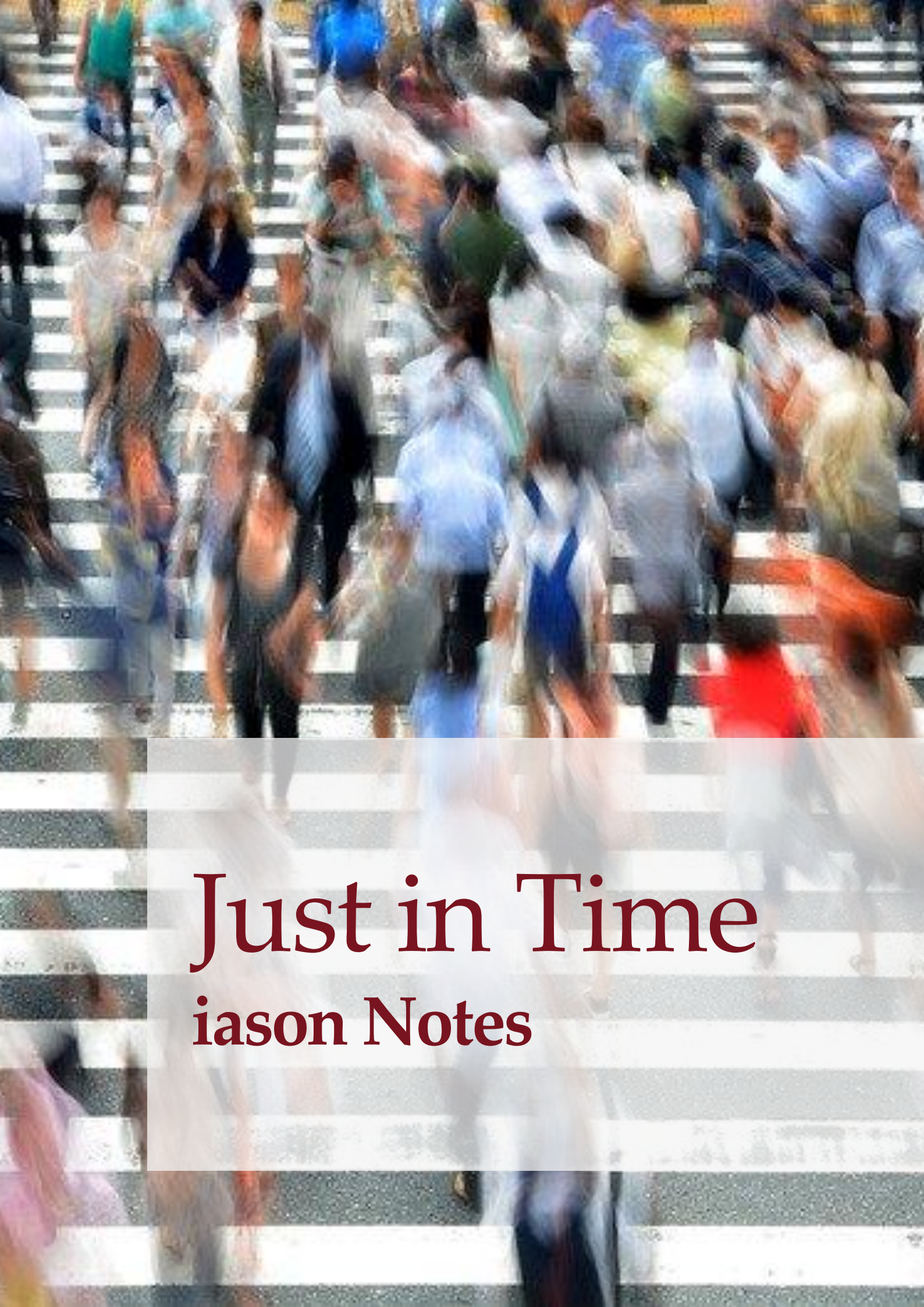
the contribution provides valuable insights into the governance challenges that institutions may face as AI systems become increasingly autonomous.

*The section concludes with “**Classifying Financial Transaction for Invoice Matching**” by M. Zanoni and E. Veksin, which addresses the automation of transaction classification to support invoice reconciliation processes. By comparing Machine Learning techniques based on k-Nearest Neighbors with Large Language Model approaches leveraging BERT architectures, the authors demonstrate how advanced AI solutions can achieve high levels of accuracy while improving scalability and operational efficiency. The article also introduces confidence-based mechanisms to identify uncertain cases requiring human review, illustrating how AI can effectively support financial control activities and allowing professionals to focus on higher-value analytical tasks.*

We hope that the articles featured in this Summer Edition provide valuable insights into the regulatory, operational, and technological developments shaping the future of financial services. As always, we invite you to engage with these topics, share your perspectives, and continue the dialogue with us.

We wish you an enjoyable reading.

Antonio Castagna
Luca Olivo



Just in Time

Jason Notes

Wealth Management



AI is emerging as a strategic driver in wealth management, improving research, client servicing, and operational efficiency across the advisory value chain. It is enabling a more data-driven and scalable model for financial services. The operating model is shifting from fully human-led to AI-augmented, with AI handling repetitive and analytical tasks. Advisors remain central for judgment, trust, and final decision-making. Strong governance and regulatory control remain essential.

[read more](#)

Date July 2026

Sound Practices for Responsible Adoption of Artificial Intelligence



The Financial Stability Board (FSB) issued this consultation report to establish a globally consistent set of sound practices for the responsible adoption of Artificial Intelligence (AI) by financial institutions. The initiative responds to the rapid acceleration of AI deployment across the financial sector, particularly following recent advances in Generative AI (GenAI) and Agentic AI, which have significantly expanded the capabilities of machine learning systems.

[read more](#)

Date July 2026

Digital Omnibus: AI Act



The Digital Omnibus constitutes a targeted regulatory initiative intended to simplify the implementation, by the other things, of the AI Act within the broader EU digital framework, without modifying its fundamental structure or risk-based logic. The proposed adjustments primarily concern implementation timelines, conformity assessment procedures, documentation requirements, relief measures for SMEs and small mid-caps, and a more structured supervisory role for the AI Office at EU level.

[read more](#)

Date June 2026

Wealth Management & Pricing Revolution: Towards More Transparent and Subscription-based Models



Wealth management pricing is undergoing a structural transformation, moving beyond traditional asset-based models under the combined pressure of regulation and market dynamics. Historically, the industry has been defined by the distinction between Assets under Management (AUM) and execution-only services, but the introduction of MiFID II has strengthened the role of advisory services (AUA), increased cost transparency and reshaping the advisor-client relationship.

[read more](#)

Date June 2026

In Data We Trust? Emerging Policy and Supervisory Approaches to AI Data Use in Financial Services



The paper examines the centrality of data in AI systems within financial services, highlighting how the rapid adoption of advanced AI - particularly generative AI (gen AI) - has amplified both the importance and risks of data governance. Data are foundational across the entire AI lifecycle: they determine model training, influence outputs, and shape performance evaluation. However, the financial sector faces structural data challenges, including fragmentation, inconsistent quality, and reliance on legacy systems.

[read more](#)

Date April 2026

The Liquidity Coverage Ratio a Decade on: A Stocktake of the Literature



The Liquidity Coverage Ratio (LCR), introduced under Basel III, requires banks to hold enough high-quality liquid assets to withstand a 30-day stress scenario. A decade of research* shows that LCR has improved banks' short-term resilience and increased holdings of liquid assets. However, it also introduced potential reduction in lending to the real economy and possible increase in risk-taking and shifts toward regulated sectors.

[read more](#)

Date July 2026

The EU Transition to Settlement T+1 Challenges, Impacts and Readiness



The transition to T+1 settlement in the EU, with go-live in October 2027, is a structural shift in the post-trade model designed to reduce counterparty risk, improve capital efficiency, and align with global standards. Key benefits include lower settlement exposure, optimized CCP margins, faster liquidity movement, and reduced operational costs through greater automation and standardization. However, the 24-hour compressed cycle creates significant operational pressure, requiring high data quality, full automation, and strong coordination to avoid settlement fails.

[read more](#)

Date May 2026

2027 EU-Wide Stress Test



The EBA has published the draft methodology, templates and guidance for the 2027 EU-wide Stress Test. While the overall framework remains broadly consistent with the 2025 exercise, targeted enhancements have been introduced to reflect recent supervisory and regulatory developments. The main methodological evolution concerns the integration of stress test reporting with the supervisory reporting framework, with greater reliance on FINREP-based data and a reduction in bespoke data requirements. This shift improves consistency and efficiency but requires institutions to enhance data mapping, lineage, validation and reconciliation capabilities.

[read more](#)

Date July 2026

EBA Report on the 2025 Credit Risk Benchmarking Exercise



The EBA published the 2025 Credit Risk Benchmarking Exercise to assess the variability of own funds requirements generated by Internal Ratings-Based (IRB) models across European Union institutions. The exercise is mandated by Article 78 of the Capital Requirements Directive (CRD) and is based on data as of 31 December 2024, collected during 2025. The overarching objective is to monitor whether differences in risk-weighted assets (RWAs) and capital requirements reflect genuine differences in risk profiles or arise from divergent modelling practices.

[read more](#)

Date July 2026

Systematic Backtesting of Probability of Default Models with Regulatory Data



The paper examines methodologies for systematic backtesting of Probability of Default (PD) models used by EU banks under the Internal Ratings-Based (IRB) framework. It addresses regulatory requirements under the Capital Requirements Regulation (CRR), which mandate regular validation of credit risk models, combining qualitative and quantitative techniques. The paper identifies limitations in traditional backtesting methodologies since asset and serial correlations are unaccounted for.

[read more](#)

Date June 2026

The EBA Streamlines Supervisory Approvals for IRB Model Changes



The EBA Final Report on the RTS amending Commission Delegated Regulation (EU) No 529/2014 introduces targeted amendments to the framework governing the assessment of materiality for changes to IRB rating systems used for credit risk capital requirements. Under Article 143 CRR, material changes require prior Competent Authorities (CA) approval, while non-material changes are subject to notification only. The RTS refines this boundary across four areas: the qualitative criteria that render a change inherently material; the scope of ex ante versus ex post notification obligations; the rules for bundling changes when evaluating quantitative thresholds; and the criteria for assessing extensions and reductions to the range of application of rating systems.

[read more](#)

Date May 2026

Incorporating Nature Physical Risks Into Banks' Credit Risk Models and Insurers' Market Risk Models: Focus on Water-related Risks



The De Nederlandsche Bank Working Paper establishes that nature-related risks - particularly those linked to biodiversity loss and ecosystem degradation - pose material threats to macroeconomic stability and financial systems. This recognition is aligned with positions taken by institutions such as NGFS, ECB and FSB. The paper highlights a critical methodological gap: while exposure-based analyses have demonstrated that financial institutions (FIs) are highly dependent on ecosystem services, there is no widely accepted framework to translate nature shocks into financial losses.

[read more](#)

Date May 2026

EBA Report on the 2025 Market Risk (IMA) Benchmarking Exercise



This report presents the results of the EBA's 2025 Market Risk Benchmarking Exercise for institutions using the Internal Models Approach (IMA). The exercise was conducted pursuant to Article 78 of the Capital Requirements Directive (CRD) and aims to assess variability in market risk capital requirements generated by banks' internal models. The report analyses the outputs of 43 European banks across 13 jurisdictions and evaluates the consistency, comparability, and supervisory implications of model-generated market risk metrics.

[read more](#)

Date July 2026

European Commission: Draft Delegated Regulation on Market Risk Prudential Requirements (FRTB)



This draft Commission Delegated Regulation introduces temporary, targeted amendments to Regulation (EU) No 575/2013 (CRR) to address implementation challenges associated with the Fundamental Review of the Trading Book (FRTB). The measures are explicitly transitional, applying from 1 January 2027 until 31 December 2029. The regulation is driven by delays and divergences in FRTB implementation across major jurisdictions, notably the United States, which risk placing EU institutions at a competitive disadvantage. To mitigate this, the Commission leverages its mandate under Article 461a CRR to introduce temporary relief and recalibration measures while maintaining alignment with Basel standards where feasible.

[read more](#)

Date May 2026

EBA Report on Simplifying the Stacking Orders of the EU Prudential and Resolution Framework



The EBA published this report* in response to increasing concerns regarding the complexity of the European Union's prudential and resolution framework. While the post-2008 regulatory reforms have successfully strengthened the resilience, solvency, liquidity, and resolvability of EU banks, they have also produced a highly layered and complex system of capital and resolution requirements.

[read more](#)

Date July 2026

EBA Initial Margin Model Authorisation under EMIR 3.0: Supervisory Requirements and Implications for Firms



EMIR 3 introduces mandatory prior authorisation by Competent Authorities (CAs) for Initial Margin (IM) models used in non-centrally cleared OTC derivatives. The goal is to ensure that material risks are adequately captured while promoting supervisory consistency and reducing duplication of work across the EU. This requirement is implemented through the EBA Draft Regulatory Technical Standards and the EBA Draft Guidelines.

[read more](#)

Date June 2026

Revised Supervisory Guidance on Model Risk Management



The guidance provides an updated supervisory framework for model risk management (MRM) applicable primarily to banking organizations with over \$30 billion in assets. It replaces earlier guidance (SR 11-7 and SR 21-8) and reflects advancements in modeling practices and supervisory experience. The revised guidance reflects the growing complexity, scope and strategic importance of models in the banking sector. Financial institutions increasingly rely on quantitative models to enhance decision-making, risk management, operational efficiency and profitability. This trend is driven by technological advancements, competitive pressures and the expansion of data-driven methodologies.

[read more](#)

Date June 2026

ECB Report on Good Practices for Climate and Nature-related Risk Stress Testing



The ECB's report provides a comprehensive overview of the evolution of climate and environmental stress testing within the European banking sector. The report builds on the ECB's 2022 Climate Risk Stress Test (CST) and the supervisory follow-up exercises conducted between 2023 and 2025. It aims to support banks in strengthening their methodologies, data infrastructures, governance arrangements, and modelling capabilities for climate and nature-related risks. The publication is particularly relevant in light of the forthcoming EBA Guidelines on environmental scenario analysis, which will apply from January 2027.

[read more](#)

Date June 2026

Good Practices for Climate and Nature Risk Management: Observations from the ECB's Five Year Climate and Nature Risk Programme (2020-25)



The ECB published this report to present observed good practices in the management of climate and nature-related (C&N) risks across significant institutions supervised within the Single Supervisory Mechanism (SSM). The paper builds on the ECB Guide on climate-related and environmental risks issued in 2020 and reflects supervisory observations accumulated during the ECB's five-year supervisory programme conducted between 2020 and 2025.

[read more](#)

Date June 2026

ESMA Promotes Proportionate Supervision of MiFID II Sustainability Requirements



The European Securities and Markets Authority (ESMA) published a public statement summarising the results of its Common Supervisory Action (CSA) on the integration of sustainability considerations into MiFID II suitability assessments and product governance frameworks. Conducted during 2024–2025 with participation from 29 EU and EEA national competent authorities (NCAs), the exercise reviewed 245 firms, including both credit institutions and investment firms.

[read more](#)

Date June 2026

ECB - Simplifying Climate Change Adaptation for Banks in the EU



The ECB Occasional Paper provides a comprehensive assessment of the regulatory gap between legal ambition and practical implementation of climate adaptation in EU banking. It highlights the need for a more coherent, standardised and forward-looking regulatory framework, capable of integrating physical risks and adaptation measures into core prudential tools. The findings support the view that adaptation must become a central pillar of sustainable finance, alongside mitigation, to ensure the resilience of both the financial system and the broader economy..

[read more](#)

Date May 2026

iason Weekly Insights

Regulatory/Supervisory Pills



Among iason's various publications we also find the iason Pills.

With these daily Pills, iason aims to offer a summary on information, mostly, of the main regulatory and supervisory news in the banking and finance sector on both Pillar I and Pillar II risks of the Basel framework. The main purpose of these publications is to give the reader an effective, timely and brief overview of the main topics of the moment.

The author of the Iason Pills is Dario Esposito.

[read more](#)

Market View



Among iason's weekly insight you can also find the iason Market View, a weekly update on financial market by Sergio Grasso.

The author, with almost three decades of investment experience, presents an accurate analysis of market fluctuations of the week, giving a critical view of observed phenomenos and suggesting interesting correlations with the main world events.

[read more](#)

GOVERNANCE. METHODOLOGY. TECHNOLOGY.

iason is a company specialised in advanced solutions for the **Risk Management** of **Financial Institutions**.

We provide highly qualified **consulting services** in the **methodological** and **technological** fields, together with targeted support for **Data** and **Model Governance** projects in risk frameworks.

We strongly believe in **Research** because we want to guarantee our clients services and solutions that are always at the forefront of **Regulatory** and **Modelling** requirements.

ARGO MAGAZINE

Quarterly magazine on the new frontiers of Risk Management

RESEARCH PAPER SERIES

Research articles on innovative and advanced methodologies in the financial sector

JUST IN TIME

Real-time updates on regulatory changes

FOLLOW US!





**Insurance Risk Management:
Overview of Solvency
Framework and Comparative
Insights on Standard and
Internal Models**

About the Authors



Gianmarco Mori:

Head of Insurance

He holds a Bachelor in Banking and Finance and a Master degree in Statistics and Actuarial science. He obtained a specialization in Quantitative Finance at Politecnico di Milano and he had worked in a consulting firm for two year focused in Credit and Liquidity Risk Management. After these years, he moved to banking sector in Risk Management area and currently he has been working in iason as financial engineer for some of the major italian players.





Matteo Bazzucchi:

Business Analyst

He holds a Master's degree in Finance and Quantitative Methods from the University of Perugia. He has over two years of experience in risk management across interest rate, credit, and operational risk domains. He has worked on projects involving IRRBB/CSRBB modelling and validation, credit risk mitigation under CRR3 and IT solutions for operational risk.



Francesco Bertonati:

Financial Engineer

Graduated in Quantitative Finance and Insurance and holding a Master's degree in Statistics and Applied Mathematics, he worked as a Financial Engineer with the Risk Management team of one of the leading Italian insurance companies, contributing to the development of the internal model. His work focused in particular on modeling dependencies between different risk factors, preparing the related technical documentation, and assessing underwriting risks in the Life insurance sector. He is currently collaborating with the Financial Risk team of one of the major Italian banks, where he focuses on risk pricing analysis and the development of models for financial products.





Daniel Bruttomesso:

Financial Engineer

He holds an MSc in Mathematical Engineering. He collaborated as a Financial Engineer with the Risk Management team of one of the leading Italian insurance companies, contributing to the development of the internal model. His work focused in particular on modeling dependencies between risk factors, preparing the related technical documentation, and assessing underwriting risks related to Life insurance. He is currently working, within the Financial Risk team of a major Italian bank, on the risk metrics engine and the pricing analysis of financial products.



Amina Cavazzana:

Credit Risk Quant

She holds a degree in Statistics for Economics and Business and currently works as a Credit Risk Analyst. She collaborates with the Risk Management team of one of Italy's leading banking groups, contributing to the development and monitoring of Probability of Default (PD) models within the Internal Ratings-Based (IRB) framework. Her main activities include credit risk model development and calibration, performance metrics and sensitivity analyses, as well as supporting model validation and methodological documentation.



This paper examines the evolution of European insurance regulation from the basic, volume-based Solvency I framework to the sophisticated and risk-sensitive Solvency II regime. It highlights how modern prudential standards increasingly align capital requirements with insurers' actual risk profiles, supported by market-consistent valuation, advanced modelling tools, and enhanced governance and disclosure expectations. A key focus is the comparison between the Standard Formula and Internal Models used to calculate the Solvency Capital Requirement (SCR). While the Standard Formula provides a harmonised and broadly applicable approach, Internal Models allow companies to reflect their specific portfolios, capturing diversification benefits, tail behaviour, and company-specific risk drivers. The paper analyses their calibration methods, underlying assumptions, and strategic implications for capital efficiency and risk management. The discussion also considers the ongoing regulatory evolution, including recent Solvency II Review proposals aimed at refining long-term guarantee measures, strengthening proportionality, and incorporating sustainability and macroprudential risks. These developments illustrate a prudential framework that continues to adapt to a changing risk environment and to insurers' role as long-term, stabilising investors. Overall, the paper offers an integrated perspective on how solvency regulation, capital models, and supervisory expectations shape modern insurance risk management and contribute to the resilience of the European insurance sector.

This paper provides a structured overview of the evolution from the traditional Solvency I framework to the modern and risk-based Solvency II regime, exploring its quantitative and qualitative pillars, its market-consistent valuation principles, and the tools that shape today's prudential environment. Building on this foundation, the paper examines the differences and strategic implications of the two principal methods used under Solvency II to determine capital requirements: the Standard Formula and Internal Models, with particular attention to their calibration, underlying assumptions, and impact on risk management practices. The analysis also considers the ongoing regulatory evolution, including recent reviews aimed at refining long-term guarantee measures, enhancing proportionality, and incorporating emerging risks such as sustainability and macroprudential concerns. By integrating historical context, actuarial concepts, and supervisory developments. The paper aims to offer a comprehensive perspective on how solvency regulation continues to adapt to support both policyholder protection and the resilience of the European insurance market. Readers already familiar with insurance regulation may want

to proceed directly to Section "Solvency II: Framework Evolution", where the Solvency II Review is introduced.

Regulatory Framework: From Solvency I to Solvency II

Introduction to Solvency and Preparatory Studies

The concept of solvency in the insurance sector refers to the ability of an insurance company to meet its obligations toward policyholders over time. It embodies the degree of assurance that an insurer can withstand adverse fluctuations in claims experience and still honour commitments. The protection of policyholders stands at the centre of solvency regulation, ensuring that insurance contracts maintain their economic and social function. From an economic perspective, solvency represents a dynamic balance between risks and the capital available to absorb them, with a specific emphasis on the time horizon over which those risks might materialise. Its importance goes beyond the safeguard of individual policyholders, since adequate solvency levels contribute to market stability, investor confidence and the effectiveness

of prudential supervision in the financial system as a whole.

The first systematic efforts to define solvency standards in Europe emerged in the post-war years, when regulators and academics recognised the necessity of calibrating capital requirements for insurers. A landmark contribution came from the Belgian actuary Campagne, who, under the mandate of the OECD, analysed loss experience data to establish a framework for minimum solvency margins. His studies initially relied on Swiss non-life insurance data from 1945 to 1954 and assumed that expenses were a stable proportion of premiums. The empirical findings suggested that loss ratios could be modelled with a Beta distribution, while the number of claims could be represented by a Poisson distribution under assumptions of independence. Using these statistical tools, Campagne estimated the probability distribution of the combined ratio and identified a capital buffer capable of absorbing extreme but plausible adverse outcomes. At a confidence level of 99.97%, the extreme combined ratio was calculated at 125%, leading to a solvency margin equal to 25% of premiums, which was interpreted as a "safety buffer" between expected results and potential ruin.

The conceptual innovation lay in treating solvency as a probabilistic measure of ruin, aligning capital with risk through quantitative methods. This approach introduced the logic of Value at Risk to insurance long before it became standard in financial regulation. The balance between protecting policyholders and preserving insurer profitability was already evident: while a larger safety buffer enhanced security, it also constrained capital available for growth and return. Subsequent extensions of the Campagne study across eight European countries confirmed the robustness of the 25% margin, though with national variations ranging from as low as 3% in Germany to as high as 35% in France. Campagne also suggested an additional loading for

reinsurance credit risk, amounting to 2.5% of ceded premiums, highlighting the interconnection between primary insurers and reinsurers in safeguarding solvency.

These preparatory studies had clear regulatory implications. They provided the first indicative coefficients that would later influence European Solvency frameworks, starting with the so-called Solvency 0 introduced in the 1970s and then inspiring parallel developments in other jurisdictions, such as the Risk-Based Capital system in the United States and the Swiss Solvency Test. By integrating statistical modelling with supervisory policy, these early contributions laid the groundwork for a progressive alignment of capital requirements with the actual risk profile of insurance undertakings, thereby initiating the gradual transition from static solvency margins to risk-sensitive approaches.

The Regulatory Framework and Solvency I (Directive 2002)

The first attempts to establish a coherent European Solvency framework date back to the 1970s, when the First Directives on non-life and life insurance introduced the concept of a minimum solvency margin, known as Solvency 0. These rules marked a significant departure from the earlier system, in which insurers were only required to hold a minimum share capital that did not depend on either premium volumes or portfolio composition. With Solvency 0, companies were required to maintain available capital at least equal to a calculated margin, derived from premiums and claims for non-life business and from mathematical reserves and sums at risk for life business. This laid the foundation for a risk-sensitive solvency regime, albeit in a still rudimentary form.

The regulatory framework was updated with Directive 2002/12/EC [10], which represented the formalisation of Solvency I. This directive, later implemented in Italy through Legislative Decree 307/2003 [4],

introduced only limited adjustments compared with the Solvency 0 system, such as a slight increase in the minimum solvency margin and in the guarantee fund thresholds, together with stronger intervention powers for the supervisory authority. The essential structure, however, remained unchanged, preserving the principle that the available solvency margin must not fall below the required solvency margin.

The architecture of Solvency I was defined through three main elements:

- **Minimum Solvency Margin (MSM):** the required capital calculated according to fixed rules based on business volume;
- **Guaranteed Fund (GF):** equal to one third of the MSM but not less than an absolute floor depending on the lines of business exercised, ensuring that companies maintained a stable minimum of available resources;
- **Available Solvency Margin (ASM, or MSD):** defined by the admissible own funds recognised under accounting and regulatory principles.

Supervisory authorities such as ISVAP, later IVASS, were responsible for verifying the compliance of insurers with these requirements and could intervene if either the MSM or the QG was breached. The solvency condition was expressed in the inequality $ASM \geq MSM$ which became the cornerstone of the system, reflecting the principle that the resources available to an insurer had always to be at least equal to the required solvency capital.

Solvency I Calculation Formulas

Within the Solvency I regime, the calculation of the Minimum Solvency Margin (MSM) was carried out through formulas that were deliberately simple but intended to reflect the main underwriting exposures of insurers.

For non-life business the requirement was determined using two alternative bases: premiums and claims. On the premium basis, the margin was linked to the volume of gross written premiums, adjusted by a retention factor to account for reinsurance. On the claims basis, the margin was computed on the average claims burden of the past three years, again adjusted for reinsurance. The formula required the higher of the two results, thereby avoiding the possibility that an unusually favourable claims history or a temporary decline in premiums could reduce the requirement excessively. An additional surcharge of 50% was applied to liability lines other than motor, reflecting the higher volatility and tail risk associated with these classes. Formally, if P denotes premiums, S the average claims burden, and α the retention factor, the required margin could be represented schematically as:

$$MSM_{NL} = \max\{\alpha \cdot k_1 \cdot P, \alpha \cdot k_2 \cdot S\},$$

where the coefficients k_1 and k_2 were defined by the directive and the surcharge applied in case of non-motor liability business. This construction made the formula only partially sensitive to risk, since it captured volume but not diversification or reserve risk.

For life business, the calculation was based on two components. The first was 4% of the gross mathematical reserves, adjusted by a minimum retention factor of 85%, reflecting the possibility that reinsurance could reduce but not eliminate the exposure. The second was 0.3% of the sums at risk, defined as the insured capital minus the reserves, again adjusted by a minimum retention factor of 50%. In symbolic form:

$$MSM_L = 0.04 \cdot V \cdot c_1 + 0.003 \cdot CR^+ \cdot c_2,$$

where V denotes the mathematical reserves, CR^+ the positive sums at risk, $c_1 \geq 0.85$ the retention coefficient for reserves, and $c_2 \geq 0.50$ the retention coefficient for sums at risk. This design aimed to capture both

the long-term accumulation of obligations (through reserves) and the immediate exposure to mortality or similar risks (through sums at risk). A further safeguard was provided by the Guarantee Fund, equal to one third of the MSM and subject to absolute minimum thresholds expressed in euro amounts depending on the type of business authorised. This mechanism ensured that even small or newly established companies would not operate with insufficient capital. The overall solvency condition remained expressed by the inequality:

$$ASM \geq MSM,$$

where ASM represented the available solvency margin, defined by admissible own funds such as paid-up capital, reserves, and certain subordinated instruments. Only resources of sufficient permanence and loss-absorption capacity could be recognised as eligible.

Structural Limitations of Solvency I

The Solvency I regime, despite introducing a more formalised prudential framework, revealed several structural limitations that progressively reduced its effectiveness as insurance markets evolved. One of the most evident shortcomings was the absence of risk differentiation in the non-life formula. The required margin was calculated on the overall portfolio without distinguishing among lines of business and with only limited recognition of volatility differences. Reserve risk was not explicitly considered, even though it represents one of the major sources of uncertainty in non-life insurance. The fixed coefficients applied to premiums and claims were based on outdated empirical studies and failed to reflect the true heterogeneity of risk profiles across markets.

In life insurance the same rigidity emerged. The 4% factor on reserves and the 0.3% factor on sums at risk were applied uniformly, irrespective of product complexity or market conditions. Although reinsurance ad-

justments were included, they were only partial and often distorted by minimum retention thresholds, favouring proportional treaties over non-proportional ones. Moreover, negative sums at risk, such as those arising from annuity portfolios, were not considered, resulting in an underestimation of biometric risks linked to longevity.

Another fundamental limitation was the neglect of financial and market risks. Asset allocation was not captured in the calculation of the minimum solvency margin, which implicitly assumed a homogeneous and low-risk investment structure. Credit risk, liquidity risk, asset-liability mismatch and operational risk were completely absent from the framework. The system therefore provided no incentive for insurers to manage these exposures actively, nor did it penalise excessive concentration in risky assets.

Finally, the design of the formulas implied that the relative capital burden decreased as company size increased, due to concavity in the relationship between MSM and business volume. This created an implicit advantage for larger insurers without any explicit consideration of diversification or correlation effects. The overall result was a system that was simple to administer and easy to verify from an accounting perspective, but one that lacked sensitivity to the actual risk profile of insurers. These limitations highlighted the need for a more comprehensive and market-consistent approach, setting the stage for the development of Solvency II.

The New Solvency II System

The rationale for the reform of the solvency regime in Europe arose from the recognition that Solvency I, while functional as a first prudential framework, was no longer able to capture the complexity of modern insurance risks. The uniform coefficients applied to premiums, claims, and reserves proved insufficient in light of financial innovation, the growing integration of capital

markets, and the emergence of new biometric and operational risks. Furthermore, the lack of a true risk-based approach and the absence of incentives for proper asset-liability management meant that the solvency requirements under the old system could not guarantee either comparability across companies or resilience under stress scenarios. Against this background, regulators sought a framework capable of aligning capital requirements with the actual risk profile of each insurer, thereby strengthening both policyholder protection and the stability of the European insurance sector.

Three-Pillar Architecture and Market-Consistent Valuation

The architecture of Solvency II (Directive 2009/138/EC [11]) was conceived on the basis of a three-pillar framework, inspired by the Basel II structure for banks but adapted to the specificities of the insurance sector. The reform aimed to create a prudential regime that is not only quantitatively robust but also qualitatively sound and transparent, capable of reflecting the true economic position of insurance undertakings.

The innovations introduced by Solvency II transformed both methodological frameworks and operational processes. On the methodological side, the shift to market-consistent valuation replaced traditional conservative accounting principles, making insurers' balance sheets a closer representation of economic reality. Assets and liabilities are valued in line with observable market data and risk-adjusted assumptions, ensuring consistency with financial markets. Capital requirements were recalibrated to rely on probabilistic measures of risk, aligning the solvency assessment with modern risk theory.

On the operational side, the new framework embedded corporate governance, risk management, and transparency into the core of prudential regulation. This approach recognizes that solvency depends not only on capital levels but also on how risks are ac-

tively identified, measured, and controlled through daily business processes.

The system was structured around three interdependent pillars, each addressing a key dimension of prudential supervision:

- **Pillar I – Quantitative requirements**

This pillar sets the capital standards, notably the Solvency Capital Requirement (SCR) and the Minimum Capital Requirement (MCR). These are designed to capture the full spectrum of risks affecting insurers, including underwriting, market, credit, and operational risks, through probabilistic models and market-consistent valuation principles. Pillar I therefore represents the quantitative foundation of the regime, ensuring that capital levels are commensurate with the underlying risk profile.

- **Pillar II – Governance and supervisory review**

This pillar introduces qualitative requirements centred on the system of governance, risk management, and the supervisory review process (SRP). It emphasises that capital adequacy alone does not guarantee solvency; the stability of an insurer also depends on the quality of its internal controls, decision-making processes, and overall risk culture. The Own Risk and Solvency Assessment (ORSA) is a key component, requiring firms to assess their solvency position in a forward-looking and holistic manner.

- **Pillar III – Disclosure and market discipline**

Transparency is elevated to a core principle. Insurers must disclose both quantitative and qualitative information regarding their solvency position, risk exposures, and governance systems through the Solvency and Financial Condition Report (SFCR) and the Regular Supervisory

Report (RSR). These disclosures aim to strengthen market discipline and external monitoring, allowing supervisors, investors, and policyholders to form a clear view of the insurer's financial soundness.

The three-pillar architecture was intended to balance numerical rigour with supervisory judgement and market discipline, creating a holistic prudential system rather than a purely formulaic one.

A central methodological innovation within Pillar I is the market-consistent valuation of assets and liabilities. For liabilities, this required moving away from the conservative technical provisions of the past and introducing a two-part structure: the Best Estimate and the Risk Margin, which will be described in detail in Section *"Valuation of Technical Provisions"*.

The Best Estimate represents the probability-weighted present value of all future cash flows, discounted at the risk-free rate and based on realistic assumptions. The Risk Margin is an additional allowance, calculated using a cost-of-capital approach, designed to ensure that the transfer value of the liabilities would be sufficient even in the presence of non-hedgeable risks.

Together, these two components established a valuation framework that is both economically grounded and risk-sensitive, fully consistent with market principles and with the overall prudential objectives of Solvency II.

Quantitative Requirements

The quantitative requirements of Solvency II were designed to create a more risk-sensitive capital framework, capable of aligning prudential standards with the true exposure of each insurance undertaking. At the core of this system lies the Solvency Capital Requirement (SCR), calibrated as the Value-at-Risk of basic own funds over a one-year time horizon with a confidence level of 99.5%. This definition implies that an insurer holding capital equal to the SCR should, under normal conditions, have less

than a 0.5% probability of insolvency within a year. The SCR is comprehensive in scope, covering underwriting risk for both life and non-life business, market risk, counterparty default risk, and operational risk. It can be calculated through the standard formula, which applies prescribed parameters and correlation structures, or through internal models approved by supervisors, provided that these models demonstrate statistical quality, robust calibration, and integration into the company's risk management processes.

Alongside the SCR, the directive introduced the Minimum Capital Requirement (MCR), defined as a lower threshold calibrated to a Value-at-Risk at the 85% confidence level over one year. The MCR represents the minimum amount of capital that an insurer must always hold in order to continue operating. Falling below the SCR triggers supervisory intervention and the requirement to restore capital within a specified period, whereas a breach of the MCR leads to immediate and more severe supervisory action, including potential withdrawal of the license. The relationship between SCR and MCR (Section *"The Minimum Capital Requirement"*) was designed to ensure a ladder of intervention, with the SCR functioning as an early warning level and the MCR as an absolute floor, thereby safeguarding both policyholders and financial stability.

Supervisory Authorities

At the European level, the European Insurance and Occupational Pensions Authority (EIOPA) plays a central role in ensuring the consistent application of the Solvency II regime across Member States. As one of the three European Supervisory Authorities (ESAs), EIOPA's mandate is to safeguard financial stability, protect policyholders and beneficiaries, and promote supervisory convergence throughout the EU insurance market.

EIOPA fulfils these responsibilities through a broad set of tools aimed at enhancing

uniformity and quality of supervision. It issues binding Technical Standards, Guidelines, and Opinions, which complement the Solvency II Directive, Delegated Acts (Delegated Regulation (EU) 2015/35 [13]) and Omnibus (2014) [12] by providing detailed, harmonised interpretations of prudential rules. National competent authorities, such as IVASS in Italy, are required to comply with EIOPA's Guidelines, ensuring a high degree of supervisory consistency.

A core element of EIOPA's mandate is the promotion of supervisory convergence. Through peer reviews, onsite visits to supervisory authorities and participation in Colleges of Supervisors, EIOPA supports the adoption of best practices across jurisdictions. This reduces the risk of regulatory arbitrage and strengthens the Single Market for insurance services. EIOPA also develops common methodologies for risk assessments, including the design and coordination of EU-wide stress tests, which are presented in the next section.

In addition to supervisory convergence, EIOPA has a substantial role in market monitoring and risk analysis. It publishes regular reports, such as the Financial Stability Report, the Consumer Trends Report, and the Insurance Risk Dashboard, which inform policymakers, supervisors and stakeholders of emerging vulnerabilities and structural trends. EIOPA also contributes to the enhancement of disclosures by setting guidelines on public reporting and by monitoring the quality of Solvency II data submitted by national authorities.

Ultimately, EIOPA's role under Solvency II is instrumental in maintaining a stable, transparent and integrated European insurance market. By fostering harmonisation, enhancing risk sensitivity and promoting high-quality supervision, EIOPA complements the work of national authorities such as IVASS and reinforces policyholder protection across the EU.

Within the Solvency II framework, IVASS plays a pivotal role as Italy's national supervisory authority, entrusted with ensuring

the full and consistent implementation of the prudential regime across the domestic insurance and reinsurance market. IVASS acts as the competent authority for Italy, responsible for enforcing compliance with the quantitative, qualitative and disclosure requirements set out in the three Pillars of Solvency II.

From a prudential perspective, IVASS is charged with monitoring the financial soundness of supervised entities. This includes verifying the adequacy of Solvency Capital Requirement (SCR) and Minimum Capital Requirement (MCR) coverage, assessing the availability and quality of Own Funds, and ensuring that capital instruments meet the criteria defined under the Tier system. Beyond pure capital metrics, IVASS evaluates the effectiveness of insurers' governance structures, internal control systems and risk management frameworks, with particular attention to the ORSA (Own Risk and Solvency Assessment), which constitutes a key element of Pillar 2 supervision.

Regarding reporting and transparency, IVASS ensures the accuracy, completeness and timely submission of all regulatory disclosures, including the Quantitative Reporting Templates (QRT), the Regular Supervisory Report (RSR) and the Solvency and Financial Condition Report (SFCR). These instruments allow the supervisor to maintain a comprehensive and forward-looking view of the risk profile of individual undertakings and of the market as a whole.

In its supervisory role, during both off-site activities and on-site inspections, IVASS naturally focuses on the areas that can most affect an undertaking's soundness: the evolution of its risk profile, the robustness of its governance and organisational safeguards, and the alignment of day-to-day practices with regulatory expectations. Off-site monitoring helps detect early signs of deterioration or unusual trends, while on-site work enables a more structured deep dive into specific domains such as underwriting practices, reserving methodologies, governance,

asset-liability management or IT infrastructures. When these activities reveal material weaknesses or practices that fall short of requirements, IVASS activates its progressive ladder of intervention, asking undertakings to implement proportionate remedial measures and, in the most severe circumstances, resorting to more intrusive actions.

Under Solvency II, the role of IVASS evolves significantly compared with the Solvency I environment. Rather than focusing primarily on compliance with static rules and formula-based margins, the supervisory approach becomes prospective, risk-based and deeply integrated with the internal processes of undertakings. This shift requires supervisors to assess not only the numerical adequacy of capital but also the robustness of risk culture, the consistency of strategic decisions with the risk appetite, and the ability of the company to withstand adverse scenarios. As a result, the IVASS supervisory model under Solvency II strengthens policyholder protection, enhances transparency and contributes to the resilience and orderly functioning of the Italian insurance market.

Quantitative Impact Studies (CEIOPS/EIOPA)

Since its inception, the development of the Solvency II framework has relied heavily on empirical testing and data-driven policy design. To ensure that regulatory standards were appropriately calibrated, the European Insurance and Occupational Pensions Authority (EIOPA) coordinated a series of Quantitative Impact Studies (QIS) and related field exercises. These studies were conducted in cooperation with national supervisory authorities and the insurance industry to assess how proposed rules would affect firms' solvency positions, risk sensitivity, and market behavior. Through this evidence-based approach, EIOPA aimed to strike a balance between prudential soundness and economic relevance, ensuring that Solvency II could strengthen policyholder

protection and financial stability while supporting insurers' ability to provide long-term investment and protection products. Specifically, the QIS objectives are to measure the financial impact of reforms, check the calibration of regulatory parameters, assess operational feasibility, and provide a structured dialogue between supervisors and industry. Their scope typically spans a broad range of insurers and business lines, combining quantitative data templates with qualitative feedback. Methodologically, they have evolved from simple stress tests into iterative exercises involving scenario analysis, calibration adjustments, and refinements based on real-world data. The most recent example, the UK Prudential Regulation Authority's QIS conducted as part of the Solvency II review (2021-22), examined reforms to the matching adjustment, risk margin, and technical provisions. Its findings showed significant potential shifts in solvency ratios and liability valuations, underlining how QIS not only quantify numerical impacts but also highlight trade-offs between prudence, competitiveness, and proportionality in regulatory design.

The Long-Term Guarantees Assessment (LTGA), launched by EIOPA in 2013, was a major field study designed to evaluate the impact and suitability of measures intended to address volatility in insurers' balance sheets under Solvency II. The exercise focused on "anti-volatility measures" such as the matching adjustment, volatility adjustment, extrapolation of the risk-free curve, and transitional provisions, all of which were aimed at smoothing short-term market fluctuations and ensuring that the regulation reflected the long-term nature of insurance liabilities. The LTGA tested how these mechanisms would influence solvency positions, capital requirements, and the availability of long-term products, especially annuities and other guaranteed savings contracts. Results showed that while such measures could mitigate artificial volatility and reduce pro-cyclicality, they also raised

concerns about complexity, comparability across jurisdictions, and the risk of masking underlying financial weakness. Ultimately, the LTGA informed the final calibration of the Solvency II framework, balancing the need to preserve prudential soundness with the policy objective of supporting insurers' role as providers of long-term financial products and institutional investors in the broader economy.

European insurance stress tests, coordinated by EIOPA since 2011, are designed to assess the resilience of the sector to adverse but plausible shocks, such as sudden interest rate movements, sharp declines in asset values, or extreme natural catastrophe scenarios. Their objectives are not only to measure potential vulnerabilities at individual and sectoral levels, but also to enhance transparency and market discipline by publishing aggregate results, and to provide supervisors with insights for macroprudential oversight. Methodologically, stress tests combine bottom-up calculations by insurers with top-down consistency checks by EIOPA, covering both capital adequacy and liquidity aspects. In addition to these periodic exercises, the EIOPA Risk Dashboard offers a more frequent, real-time overview of risks facing the insurance sector. Based on a set of risk indicators covering solvency, profitability, market, liquidity, and macroeconomic exposures, the dashboard tracks trends across the industry and signals emerging vulnerabilities. Together, the stress tests and the Risk Dashboard provide complementary tools: the former offering deep, scenario-based assessments, and the latter ensuring continuous monitoring, both essential to the stability and supervisory effectiveness of the European insurance market.

Taken together, the QIS exercises, the LTGA, the European stress tests, and the EIOPA Risk Dashboard were fundamental in shaping the final calibration of Solvency II. QIS provided the empirical foundation for testing different capital and valuation approaches, ensuring that regulatory param-

eters were not only theoretically sound but also workable for a diverse industry. The LTGA specifically addressed the treatment of long-term guarantees, leading to the introduction and fine-tuning of anti-volatility measures such as the matching and volatility adjustments. Stress tests added a forward-looking, systemic dimension, demonstrating how insurers would cope under severe shocks and highlighting vulnerabilities that informed supervisory priorities. Finally, the Risk Dashboard established an ongoing monitoring framework, giving regulators and markets greater visibility of emerging risks. Together, these tools helped calibrate Solvency II in a way that balances prudential soundness with proportionality, supports the provision of long-term products, and strengthens the resilience and transparency of the European insurance sector.

In recent years, EIOPA has continued to deepen its empirical and supervisory work beyond classic QIS and stress-test exercises, reflecting new priorities and evolving risks in the European insurance sector. Notably, its Peer Review Work Plan 2023-2024 includes a dedicated review of stochastic valuation practices, especially in life insurers with significant options and guarantees, signaling a shift toward more sophisticated liability modeling [7]. Moreover, in 2024 EIOPA published a comparative study on diversification in internal models, highlighting large heterogeneity in how insurers capture diversification effects, a topic with direct implications for capital calibration [8]. On the risk-monitoring side, the Insurance Risk Dashboard remains a key tool: its most recent editions (e.g. April 2025) show that overall risk levels in the sector remain at medium levels, but with increasing concern around geopolitical instability, market volatility, and digital/cyber risks [6]. These newer studies complement the earlier QIS, LTGA, and stress test exercises by providing granular insights into internal model practices and continuous, real-time risk surveillance, informing regulatory cali-

bration and supervisory priorities. Together, these continuing initiatives illustrate EIOPA's commitment to maintaining Solvency II as a living framework, responsive to emerging risks and grounded in quantitative evidence.

Solvency II: Principles and Tools

The Solvency Capital Requirement (SCR)

The Solvency Capital Requirement (SCR) is the cornerstone of the Solvency II quantitative framework and represents the main solvency requirement for insurers. Its purpose is to ensure that insurance undertakings hold sufficient capital to withstand severe but plausible adverse conditions over a one-year horizon. To compute this requirement, the directive allows undertakings to choose between two primary regulatory approaches, subject to supervisory approval: the Standard Formula, which is a predefined, modular framework designed by EIOPA to capture the risk profile of an average insurer, and Internal Models (either full or partial), which are custom risk assessment frameworks developed by the company itself to better reflect its specific and complex risk characteristics.

Under the standard framework, the SCR has a modular structure (see Figure 5), composed of the Basic SCR (BSCR), operational risk, and adjustments reflecting the loss-absorbing capacity of technical provisions and deferred taxes. The general aggregation formula is:

$$SCR = \sqrt{\sum_i \sum_j SCR_i \cdot \rho_{ij} \cdot SCR_j} + SCR_{oper} - Adj,$$

where:

- SCR_i and SCR_j represent the capital requirements of risk modules and sub-modules;

- ρ_{ij} are correlation coefficients defined by the directive;
- SCR_{oper} is the capital requirement for operational risk;
- Adj reflects the loss-absorbing capacity of technical provisions and deferred taxes.

The risk modules considered in the BSCR are:

- **Market risk**, covering:
 - Interest rate risk;
 - Equity risk;
 - Spread risk;
 - Currency risk;
 - Concentration risk;
 - Property risk.
- **Counterparty default risk**, addressing exposures to reinsurers, intermediaries, and other counterparties.
- **Life underwriting risk**, including:
 - Mortality;
 - Longevity;
 - Disability-morbidity;
 - Lapse;
 - Expense;
 - Revision;
 - Catastrophe risks.
- **Non-life underwriting risk**, covering:
 - Premium risk;
 - Reserve risk;
 - Lapse risk;
 - Catastrophe risks.
- **Health underwriting risk**, addressing analogous risks for health insurance products.

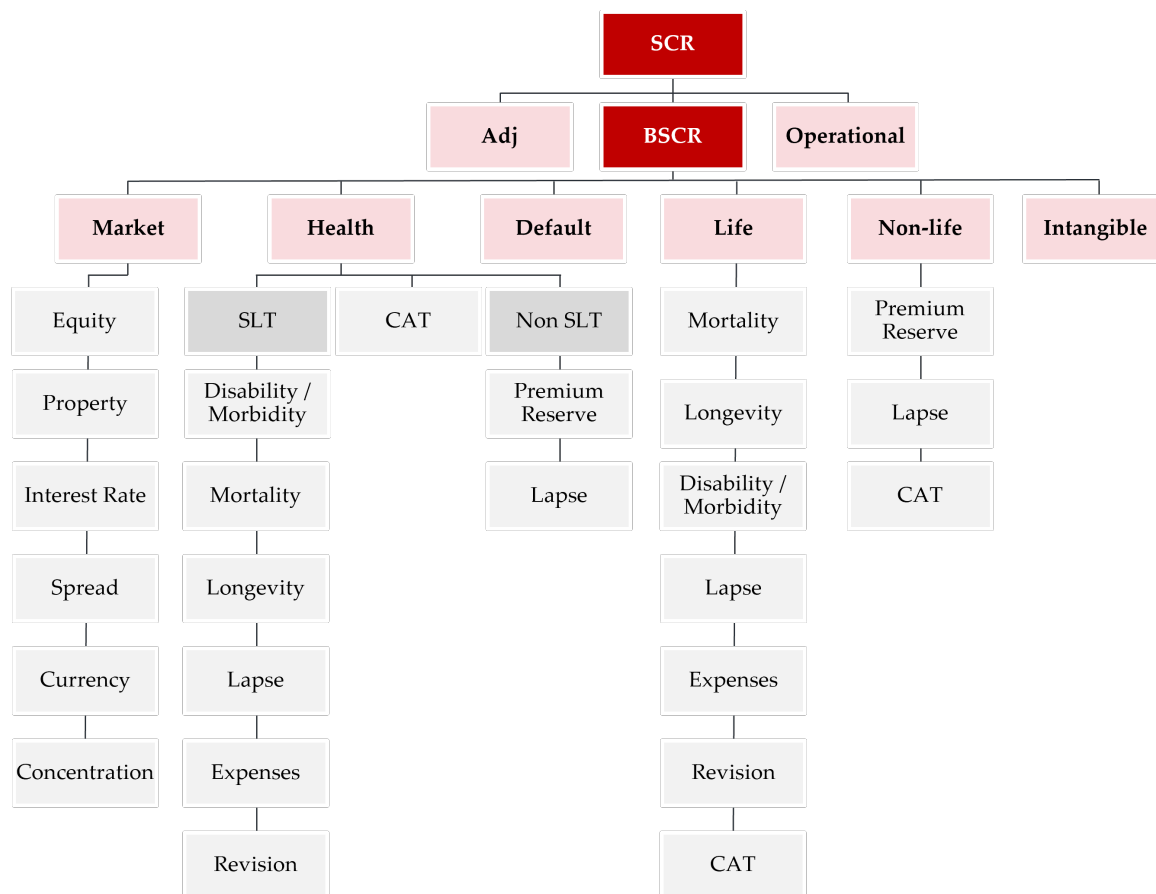


FIGURE 1: Risk aggregation tree under the Standard Formula from "Study on Diversification in Internal Models" (EIOPA) [8].

The Basic Solvency Capital Requirement (BSCR) is obtained by aggregating the individual risk modules using a correlation matrix that captures the statistical dependencies between them. This approach recognises that not all risks materialise simultaneously or perfectly correlate, thereby introducing diversification effects that reduce the overall capital requirement. However, these effects are limited by regulatory constraints to avoid excessive capital relief. The use of a linear correlation framework assumes elliptical joint distributions and may underestimate extreme co-movements, particularly during market stress.

To address this limitation, internal model approaches often employ copulas or non-linear dependence structures, providing a more realistic representation of tail correlations between risk drivers such as market and underwriting risk. Alternative approaches include multi-factor models,

where dependencies arise through common underlying risk drivers, and scenario-based frameworks, which reflect joint risk materialisation through coherent adverse economic and insurance stress scenarios. Each module is subdivided into sub-modules, calibrated individually and then aggregated through an appropriate dependency structure. This structure introduces diversification effects in a controlled way: companies can benefit from the reduction of overall capital requirements when risks are imperfectly correlated, but only within the bounds defined by the regulation.

The Solvency Capital Requirement (SCR) is defined as the level of own funds corresponding to the Value-at-Risk (VaR) of basic own funds over a one-year time horizon with a confidence level of 99.5%. This calibration means that an insurer holding capital equal to the SCR should have a probability of insolvency of less than 0.5% within

a year. Formally:

$$SCR = VaR_{99.5\%}(\Delta \text{Own Funds}, 1 \text{ year}),$$

where the variation in own funds ($\Delta \text{Own Funds}$) is driven by market, credit, underwriting, and operational risks.

Value-at-Risk (VaR)

The Value-at-Risk (VaR) is a quantile-based statistical measure that estimates the maximum loss an institution may experience over a given time horizon, at a specified confidence level, under normal market conditions. Formally, for a random variable L representing losses, the Value-at-Risk at confidence level α is defined as:

$$VaR_{\alpha}(L) = \inf\{l \in \mathbb{R} \mid \mathbb{P}(L \leq l) \geq \alpha\}.$$

Intuitively, $VaR_{99.5\%}(L)$ corresponds to the loss level that will not be exceeded with 99.5% probability over the chosen time horizon. Thus, only 0.5% of outcomes are expected to result in a greater loss. The rationale behind using VaR in the Solvency II framework lies in its clarity and interpretability: it provides a single, easily communicable number representing the capital amount needed to cover unexpected losses up to a given confidence level. Moreover, VaR is directly compatible with the balance sheet structure, as it measures potential variations in own funds based on the joint distribution of underlying risks.

Tail-Value-at-Risk (TVaR)

While the Value-at-Risk (VaR) at the 99.5% confidence level represents the official risk measure under Solvency II, it differs conceptually from the Tail-Value-at-Risk (TVaR), also known as Expected Shortfall (ES). The TVaR captures the expected loss beyond the VaR threshold, providing a more conservative and comprehensive measure of tail risk. Formally, it can be expressed as:

$$TVaR_{\alpha}(L) = \mathbb{E}[L \mid L > VaR_{\alpha}(L)].$$

Although both are percentile-based measures, TVaR averages the extreme outcomes in the tail of the loss distribution, rather than setting a fixed quantile cut-off.

Capital-at-Risk (CaR)

A complementary interpretation of the regulatory capital measure comes from the notion of Capital-at-Risk (CaR), widely used in actuarial risk theory. While the Solvency II Directive defines the SCR as the 99.5% Value-at-Risk of the one-year change in own funds, the CaR expresses the same concept by decomposing total losses into their expected and unexpected components.

Let L denote the random variable describing one-year losses. Then the Capital-at-Risk is defined as:

$$CaR_{\alpha}(L) = VaR_{\alpha}(L) - \mathbb{E}[L].$$

In this formulation, the CaR isolates the portion of losses corresponding to unexpected deviations from the mean, i.e. the economic capital required to protect the insurer against adverse outcomes occurring with probability $1-\alpha$. At the 99.5% confidence level used for the SCR, the CaR therefore represents the capital buffer needed to absorb extreme but plausible losses, consistent with the “one-in-200-year” event underlying Solvency II calibration. Accordingly, the CaR reinforces the idea that the SCR is not a measure of average performance but a quantile-based capital buffer designed to ensure policyholder protection under severe but credible stress scenarios.

Comparison with Banking Regulation

Solvency II favours the VaR approach due to its interpretability and computational simplicity, as well as its consistency with accounting data and reporting requirements. In contrast, the Basel II and Basel III frameworks for banking supervision often rely on Expected Shortfall (ES), which is sub-additive and therefore more coherent for aggregating risks in portfolios with fat-tailed

loss distributions. The choice of VaR in Solvency II ultimately reflects a compromise between statistical rigor and practical implementation, ensuring consistency and comparability across insurers while maintaining computational tractability.

Solvency Capital Requirement in the CRM

The calculation of the Solvency Capital Requirement (SCR) is grounded in the actuarial theory of collective risk, known as the Collective Risk Model (CRM).

In this framework, the insurance risk is represented by a random variable describing the aggregate cost of claims over a specified time horizon (typically one year). This cost is defined as the sum of individual claim amounts $Z_1, Z_2, \dots, Z_N$, where N is a random variable denoting the number of claims, generally assumed to follow a Poisson distribution or a Mixed Poisson distribution to account for portfolio heterogeneity.

An important generalisation of the Poisson model is the Negative Binomial distribution, which arises when the Poisson parameter itself is considered a random variable following a Gamma distribution. This mixture captures the overdispersion typically observed in real insurance portfolios, where the variance of the number of claims exceeds the mean due to portfolio heterogeneity, contagion effects, or varying exposure levels.

By allowing for additional variance, the Negative Binomial model provides a more flexible and realistic representation of claim frequency and improves the calibration of the risk modules under Solvency II. It is especially useful for large, diversified portfolios where risk pooling does not completely eliminate volatility in claim counts.

Modeling risk through the CRM allows for a consistent estimation of the mean, variance, and skewness of the loss distribution, parameters that are essential for calibrating the risk modules within the Standard For-

mula or Internal Models. A key feature of this framework is the distinction between frequency and severity of claims:

- Frequency is described by discrete distributions (e.g., Poisson, Negative Binomial);
- Severity by continuous positive distributions (e.g., Lognormal, Gamma, Pareto).

The choice of the severity distribution plays a crucial role in determining the overall variability of the aggregate claims cost. Distributions such as the Lognormal and Pareto exhibit heavy tails, meaning that a small proportion of claims can contribute disproportionately to total losses. This feature has a direct impact on the tail of the aggregate loss distribution and therefore on the Solvency Capital Requirement, which is calibrated to extreme quantiles. Selecting an appropriate severity model, supported by empirical data and statistical fitting, is therefore essential for realistic capital estimation and effective risk management.

The aggregation of the two components, claim frequency and claim severity, produces the total loss distribution, enabling the estimation of the Value-at-Risk at a 99.5% confidence level, as required under Solvency II. This approach bridges risk-management practice with statistical and actuarial theory, providing a quantitative foundation for determining both economic and regulatory capital.

Since the exact convolution of frequency and severity distributions is analytically complex, approximation methods are commonly used to estimate the distribution of aggregate losses. Classical approaches include the Normal approximation (based on the Central Limit Theorem) and refinements such as the Normal-Power and Wilson-Hilferty transformations, which improve accuracy for skewed and positively bounded data. These methods allow for the practical computation of percentiles and Value-at-Risk estimates without requiring extensive simulation.

When loss distributions are strongly asymmetric or heavy-tailed, more robust numerical techniques are recommended. The Panjer recursion formula provides a recursive algorithm to determine the probability distribution of total losses when both claim frequency and severity follow specific classes of distributions (such as the $(a, b, 0)$ class). This method is particularly suitable for portfolios with discrete claim amounts and is frequently used in actuarial practice for its speed and numerical stability.

Alternatively, Monte Carlo simulations offer a flexible approach for modelling more complex or heavy-tailed loss distributions. By repeatedly sampling both the number and the size of claims, these simulations generate an empirical loss distribution from which quantiles such as the Value-at-Risk or Tail-Value-at-Risk can be estimated directly. While computationally intensive, Monte Carlo techniques allow actuaries to incorporate dependencies, reinsurance structures, and non-linear effects that cannot be captured by closed-form formulas, providing a robust foundation for the calculation of the Solvency Capital Requirement. A more detailed discussion of analytical approximations for compound frequency-severity processes, including the Single Loss Approximation developed in the Operational Risk literature, is provided in Section “*Advanced Modelling Techniques*”.

Technical and Actuarial Insights into Non-life Underwriting Risk

Within the Solvency II framework, two key components of the non-life underwriting module are the Premium Risk and the Reserve Risk, both of which capture the uncertainty in future claims development over a one-year horizon.

From a mathematical perspective, Premium Risk relates to claims that will arise from policies not yet fully earned at the valuation date. It is typically modelled as the variability of the aggregate loss ratio over the forthcoming year. Denoting the expected

number of claims by λ and the average claim size by c_z , the total claims amount can be expressed as $S = \sum_{i=1}^N Z_i$, with $N \sim \text{Poisson}(\lambda)$ and Z_i representing independent and identically distributed claim severities. The volatility of this risk arises from both the frequency ($\text{Var}[N]$) and the severity ($\text{Var}[Z]$) components. The Reserve Risk, by contrast, reflects the potential for deviation between the best estimate of outstanding claims and their actual ultimate cost. It is often assessed through a one-year re-reserving model, in which the uncertainty of prior accident years’ liabilities is re-evaluated after one year. This dynamic perspective aligns the actuarial process with the Solvency II one-year horizon assumption for the SCR.

In stochastic formulations, the variability of Premium and Reserve Risk is parameterised through the expected frequency λ , the mean claim cost c_z , and the associated variances. The total variance of claims can be decomposed as:

$$\text{Var}(S) = \mathbb{E}[N] \cdot \text{Var}(Z) + \text{Var}(N) \cdot (\mathbb{E}[Z])^2.$$

This relationship highlights how uncertainty in both the number and size of claims contributes to overall volatility. Within the Solvency II standard formula, these risks are captured through volume measures (premiums and provisions) scaled by standard deviations calibrated by EIOPA to reflect the 99.5% Value-at-Risk confidence level. In internal model frameworks, actuaries often estimate these variances empirically, incorporating dependencies across lines of business via copulas or correlation matrices. The resulting capital requirement for underwriting risk is thus directly linked to the stochastic structure of N and Z , ensuring consistency between empirical observation and regulatory calibration.

A complementary source of uncertainty within the underwriting module is the volatility of the expense ratio, which influences both Premium and Reserve Risk. Fluctuations in administrative and acqui-

sition expenses affect the total cost structure and therefore the insurer's combined ratio. Empirical studies show that variations in expenses can amplify the overall loss ratio, particularly when expense assumptions are based on fixed overheads rather than scalable costs. Under Solvency II, the interaction between loss ratio and expense volatility is reflected in the calibration of the standard deviation parameters for Premium Risk. Higher variability in expense ratios translates into increased volatility of technical results and therefore higher capital charges. Proper modelling of expense dynamics, especially when using internal models, is crucial for accurately assessing the Solvency Capital Requirement and avoiding systematic underestimation of underwriting volatility.

The Minimum Capital Requirement

The Minimum Capital Requirement (MCR) represents the absolute floor of the Solvency system. Its function is to ensure that insurance undertakings maintain a basic amount of own funds below which they cannot legally continue operating. While the SCR provides a comprehensive and risk-sensitive benchmark, the MCR is intentionally simpler, relying on linear formulas applied to volume measures such as written premiums, technical provisions, and sums at risk.

Formally, the directive defines the MCR as a linear function:

$$MCR = f(P, V, SR),$$

where:

- P = volume of premiums;
- V = technical provisions;
- SR = sum at risk.

The function f is specified by regulatory coefficients calibrated to approximate a Value-at-Risk (VaR) at the 85% confidence level over a one-year horizon. This calibration implies that an insurer holding capital equal

to the MCR should face no more than a 15% probability of default within twelve months.

To maintain proportionality between the two capital thresholds, the directive sets explicit bounds that tie the MCR to the SCR:

$$0.25 \cdot SCR \leq MCR \leq 0.45 \cdot SCR.$$

This corridor prevents the MCR from being disproportionately low or excessively high relative to the SCR. The supervisory ladder of intervention derives from this two-tiered structure:

- If the Available Solvency Margin (ASM) is greater than or equal to the SCR, the insurer is considered fully compliant;
- If the ASM falls below the SCR but remains above the MCR, the company must present a recovery plan and restore its solvency within a prescribed period;
- If the ASM drops below the MCR, the situation is deemed critical and immediate supervisory action follows, potentially leading to the withdrawal of authorisation.

Through this combination of linear calculation, probabilistic calibration, and regulatory boundaries, the MCR provides a simple but decisive safeguard against insolvency, complementing the broader solvency assessment offered by the SCR.

Valuation of Technical Provisions

A central innovation of Solvency II is the introduction of a market-consistent valuation of technical provisions, replacing the static and prudential methods of Solvency I. The objective is to align the balance sheet of insurers with principles consistent with financial markets, ensuring that liabilities are measured at a realistic transfer value. The value of technical provisions is defined as the sum of two components: the Best

Estimate (BE) and the Risk Margin (RM).
 $TP = BE + RM.$

Best Estimate

From an actuarial perspective, the Best Estimate represents the expected present value of future losses arising from insurance obligations. Its assessment is based on stochastic models such as the Collective Risk Model (CRM), which describes the distribution of the aggregate cost of claims and allows for the estimation of key moments (mean, variance, and skewness). These parameters ensure that the reserve accurately reflects the true risk profile of the company, in line with the market-consistent valuation approach required by Solvency II.

In operational terms, the Best Estimate (BE) is calculated as the present value of all expected future cash flows arising from insurance liabilities, split between premium provisions (relating to unexpired risks) and claims provisions (relating to incurred but unpaid losses). For non-life business, the premium provision captures expected claims and expenses during the remaining coverage period, while the claims provision represents the discounted value of outstanding claims, including incurred but not reported (IBNR) losses. Actuaries often rely on the combined ratio (the sum of the loss ratio and the expense ratio) as an empirical indicator for assessing profitability and estimating expected future cash flows. By projecting future combined ratios under realistic assumptions of frequency, severity, and expenses, insurers can derive a robust estimation of the Best Estimate that aligns with the company's actual risk experience. The Best Estimate corresponds to the probability-weighted average of all future cash flows arising from insurance obligations, discounted at the risk-free interest rate term structure. Formally:

$$BE = \sum_{t=1}^T \frac{\mathbb{E}[CF_t]}{(1 + r_t)^t},$$

where $\mathbb{E}[CF_t]$ denotes the expected cash

flow at time t and r_t the relevant risk-free rate for maturity t . This removes the implicit prudence margins typical of traditional reserving and ensures that liabilities reflect realistic expectations.

Risk Margin

For the Risk Margin component, the actuarial approach adopts the Cost-of-Capital method, applied to the capital required to cover non-hedgeable risks over the run-off period of the liabilities. Within this framework, the CRM provides a useful structure for estimating tail risks, which play a critical role in determining the magnitude of the Risk Margin. Monte Carlo simulation techniques or analytical approximations (such as the Normal-Power or Wilson-Hilferty methods) are frequently used to estimate the extreme percentiles of the loss distribution, thereby enhancing the stability and robustness of actuarial valuations.

The Risk Margin (RM) is determined using the Cost-of-Capital (CoC) approach, which reflects the cost of maintaining sufficient capital to support non-hedgeable risks over the lifetime of the obligations. The regulatory CoC rate is currently set at 6% per annum in the Solvency II framework, applied to the projected Solvency Capital Requirements for non-hedgeable risks ($SCR_{NH,t}$) at each future time period t . Formally, the Risk Margin is given by:

$$RM = \sum_{t=1}^T \frac{CoC \cdot SCR_{NH,t}}{(1 + r_t)^t}.$$

This ensures that, if the liabilities were transferred to another insurer, sufficient funds would be available to remunerate the capital required to bear these residual risks. In practice, insurers often approximate the projection of $SCR_{NH,t}$ using simplifications based on proportional decay or duration-adjusted factors, provided these remain consistent with the underlying risk profile.

Risk-free Interest Rate Term Structure

The risk-free interest rate term structure used for discounting technical provisions is published regularly by EIOPA and is based on euro swap rates adjusted for credit risk. For long maturities where observable market data are limited, the curve is extrapolated to an Ultimate Forward Rate (UFR), currently set at 3.25% per annum for the euro.

Two major adjustments can be applied to the risk-free curve:

- The Volatility Adjustment (VA), designed to mitigate artificial fluctuations in solvency ratios caused by temporary market volatility;
- The Matching Adjustment (MA), applicable to portfolios with predictable cash flows and eligible backing assets, allowing the discount rate to better reflect asset yields.

Both measures aim to ensure that the valuation of technical provisions reflects the long-term economic reality of insurance contracts, balancing prudence with stability.

The removal of implicit prudence margins and the consistent treatment of hedgeable and non-hedgeable risks represent a fundamental departure from previous frameworks. Assets and liabilities are now valued under the same principles, enhancing comparability, while the explicit split between Best Estimate and Risk Margin increases transparency in the solvency position.

The Solvency II Balance Sheet

The Solvency II balance sheet represents a fundamental shift from the accounting-based approach of Solvency I to an economic and market-consistent framework. Its purpose is to provide a realistic picture of the insurer's financial position by valuing both assets and liabilities on a basis consistent with observable market conditions.

On the liability side, technical provisions are calculated as the sum of Best Estimate and Risk Margin, ensuring that obligations reflect expected future cash flows and the cost of capital needed to cover non-hedgeable risks. On the asset side, valuations are made at market-consistent prices, removing historical cost accounting and embedding real-time information into solvency measurement. This consistency between assets and liabilities allows for a meaningful calculation of the insurer's available capital.

The measure of available capital under Solvency II is represented by Own Funds, defined as the excess of assets over liabilities according to the Solvency II valuation rules:

$$\text{Own Funds} = \text{Assets} - (\text{BE} + \text{RM}).$$

Own Funds are classified into three quality categories, or Tiers, based on their permanence, subordination, and loss-absorbing capacity:

- Tier 1: the highest-quality capital, including paid-up ordinary share capital, retained earnings, and other fully available reserves. These funds are permanent and fully loss-absorbing.
- Tier 2: subordinated liabilities and hybrid instruments with limited permanence and availability, which can absorb losses but to a lesser extent than Tier 1.
- Tier 3: the lowest-quality category, including items such as deferred tax assets, which have a limited ability to absorb losses and are often subject to stricter quantitative restrictions.

Eligibility limits are imposed on the use of Own Funds to cover capital requirements, in order to ensure that solvency is supported primarily by the most reliable resources. In particular, coverage of the Minimum Capital Requirement (MCR) must be met exclusively with Tier 1 and Tier 2 items, with strict caps on the latter, while the Solvency Capital Requirement (SCR) may be

covered with a broader mix of Tiers but still subject to eligibility ratios.

This balance sheet design emphasises economic substance over accounting form, ensures comparability across jurisdictions, and provides supervisors and markets with a transparent view of the solvency position of each insurer. It is thus the backbone of the Solvency II regime, integrating valuation rules, technical provisions, and capital classification into a single coherent framework.

Solvency II: Framework Evolution

The evolution of the Solvency II framework shows that European insurance regulation is not meant to be a fixed or static regime but a prudential system that adapts to changing market conditions and emerging macroeconomic challenges. Solvency II has gradually incorporated new tools, recalibrated quantitative requirements and expanded qualitative expectations in response to shifts in the risk environment, from prolonged low interest rates to the shocks caused by the COVID-19 pandemic and the effects of geopolitical instability. Each revision, whether through delegated regulations or broader legislative amendments, aims to preserve the core objective of protecting policyholders while ensuring that insurers can continue to operate as long-term investors, contributors to financial stability and providers of private capital to the real economy. In particular, the updates introduced by Directives 2025/2 therefore demonstrate the built-in flexibility of the Solvency II framework, which evolves alongside market practice, supervisory experience and the Union's strategic priorities including sustainability, macroprudential resilience and support for long-term investment.

Directive (EU) 2025/2 - Solvency II Review

Directive (EU) 2025/2 [14] is adopted against the backdrop of a sequence of severe shocks to the Union's economy. The COVID-19 pandemic generated substantial social and economic damage, exposing structural vulnerabilities and calling for a recovery strategy that is not merely rapid but also sustainable, inclusive and fair. The economic and social consequences of Russia's war of aggression against Ukraine have added further pressure on the Union's policy priorities, for example through energy price volatility, heightened macroeconomic uncertainty and renewed concerns about financial stability. In this context, the insurance and reinsurance sector is identified as a key channel through which private capital can be mobilised to support the real economy and to strengthen the resilience of European businesses.

The revision of the Solvency II framework through Directive (EU) 2025/2 pursues several interrelated objectives. A central aim is to encourage long-term investments by institutional investors, in particular insurance undertakings, so as to support the re-equity of the corporate sector and to facilitate the financing of the Union's economic recovery. To this end, the prudential framework is adjusted to better reflect the structurally long-term nature of insurance liabilities and investment horizons. Among other elements, the Directive facilitates the use of a more favourable standard parameter for equity investments held with a genuinely long-term perspective, on condition that the undertaking can demonstrate its capacity to avoid forced sales in stressed market conditions. The capital freed through the reduction of the Solvency Capital Requirement (SCR) is expected to be channelled primarily towards productive investments in the real economy rather than being used for shareholder distributions or variable remuneration.

At the same time, the Directive aligns Solvency II more closely with the objectives

of the European Green Deal. It seeks to transform the Union into a climate-neutral and climate-resilient economy by 2050, in part by integrating climate and environmental risks more coherently into the prudential framework. Insurance undertakings are required to embed sustainability risks in their risk-management systems and to conduct climate-related scenario analyses over long-term horizons. The reform also reinforces the principle of proportionality so that small and less complex undertakings are not unduly burdened, and it strengthens supervisory tools and cooperation, especially with regard to cross-border activities and the monitoring of systemic risk in the insurance sector.

Proportionality and Scope of Application

A core dimension of Directive (EU) 2025/2 is the recalibration of proportionality and the refinement of the scope of application of the Solvency II regime. The legislator recognises that the framework should not apply inappropriately to very small undertakings and that the intensity of requirements must be aligned with the nature, scale and complexity of the risks inherent in each undertaking's business.

At the level of full exclusion from the regime, the quantitative thresholds determining when an undertaking falls outside Solvency II are increased. An undertaking may be excluded where its annual gross written premium income does not exceed EUR 15,000,000 and the total amount of its technical provisions, gross of recoverables, does not exceed EUR 50,000,000. For undertakings that belong to a group, the group's technical provisions must also remain below this threshold. Undertakings meeting these conditions may operate entirely outside Solvency II, while retaining the option to request authorisation in order to benefit from the single-licence framework. The conditions for excluding assistance activities are also revised, allowing occasional provision of roadside-assistance services in

neighbouring Member States without altering the regulatory treatment of the undertaking.

Above the exclusion threshold, the Directive establishes a structured proportional regime for "small and non-complex undertakings" (SNCU) [9]. These are identified through a set of risk-based criteria that differ depending on whether the undertaking conducts life, non-life or mixed business, and the criteria must be satisfied for two consecutive years. For example, in the non-life sector, annual gross written premiums may not exceed EUR 100,000,000 and cross-border premium income may not exceed EUR 20,000,000 or 10% of total premiums. In the life sector, one criterion requires that the interest-rate risk exposure remain limited relative to the undertaking's technical provisions.

Classification as an SNCU grants automatic access to a broad set of proportionality measures relating to reporting, disclosure, governance, the own-risk and solvency assessment (ORSA) and liquidity-risk management. SNCU are exempted from the requirement to prepare a formal liquidity-risk management plan. The classification process is intentionally streamlined: the undertaking notifies its supervisory authority and, in the absence of objection within a defined period, acquires SNCU status.

The Directive also addresses the position of captive undertakings. Captives may be classified as SNCU even where they do not meet all general criteria, provided specific conditions are satisfied, such as limiting policyholders and beneficiaries to entities within the same group and avoiding coverage of mandatory third-party liability risks. Supervisory authorities maintain discretion to restrict an SNCU from using proportionality measures if serious concerns arise regarding its risk profile or governance. Conversely, undertakings that do not qualify as SNCU may still be permitted, following supervisory assessment, to rely on certain proportionality measures.

The resulting design distinguishes clearly

between full exclusion from Solvency II for very small undertakings and SNCU status for undertakings that remain within the system but benefit from proportionate and simplified regulatory obligations.

Governance, Risk Management and Sustainability

Directive (EU) 2025/2 reinforces the qualitative dimension of the Solvency II regime (Pillar II) by strengthening governance requirements and broadening the scope of risk management to include macroprudential and sustainability-related considerations. Supervisory authorities are required to conduct periodic reviews of undertakings' governance systems as an integral part of the supervisory review process, assessing both the formal arrangements and their effective implementation in practice.

Persons who effectively run the undertaking or perform key functions must satisfy fit-and-proper requirements on a continuous basis in terms of competence, experience and integrity. Where these conditions are no longer met, the supervisory authority has the power to require the removal of the person concerned. Undertakings must adopt a policy promoting diversity at the level of the administrative, management or supervisory body, including the specification of quantitative objectives related to gender balance. The Directive also clarifies the expectations for key functions: risk management, actuarial, compliance and internal audit functions must be organised so as to ensure independence and effective challenge. For SNCU, certain combinations of key functions are allowed under defined conditions, with the important exception of the internal audit function, which must remain separate. The integration of environmental, social and governance (ESG) risks is a prominent feature of the reform. Undertakings must establish strategies, policies, processes and systems for the identification, measurement, management and monitoring of sustainability risks over short,

medium and long-term horizons. They are required to develop and monitor the implementation of specific plans that include quantifiable targets for addressing financial risks stemming from sustainability factors, in line with relevant Union climate and environmental objectives. These plans are expected to influence underwriting practices, investment strategies and product design, thereby embedding sustainability considerations in the core business model of insurance undertakings.

The own-risk and solvency assessment is significantly expanded. Undertakings with material exposure to climate-related risks must assess the impact of at least two long-term climate scenarios on their risk profile and solvency position. One scenario must be consistent with a global temperature increase below two degrees, while another must reflect a significantly higher increase, capturing transition and physical risks under different policy and technological pathways. These scenario analyses must be carried out at regular intervals. SNCU are exempted from this specific requirement in recognition of their more limited resources and risk exposures. Beyond climate, the ORSA must systematically address macroprudential aspects: undertakings have to consider macroeconomic conditions, potential developments in financial markets and, where requested by the supervisory authority, structural macroprudential concerns such as credit cycles, herding behaviour in investment strategies or excessive concentrations at market level.

This macroprudential perspective must also be incorporated into underwriting and investment activities and into risk-management practices more broadly. Undertakings are expected to take into account the collective behaviour of market participants and the potential amplification of risks through common exposures or correlated strategies. Supervisory authorities, in turn, are required to share their assessments with relevant national bodies, thereby linking microprudential supervi-

sion with the broader framework of financial stability oversight. Cross-border activities considered significant receive particular attention, with formalised information flows between home and host authorities and the possibility for host supervisors to address requests directly to undertakings or to participate in joint on-site inspections where there are indications of non-compliance with capital requirements.

Quantitative Requirements, Macroprudential Tools, Supervision and Reporting

On the quantitative side (Pillar I), Directive (EU) 2025/2 introduces important changes to the valuation of technical provisions, to the calibration of the risk margin and to elements of the standard formula for the SCR. These changes are closely linked to the objective of supporting long-term business models while preserving a high level of policyholder protection.

The calculation of the risk margin is modified to take explicit account of the time dependency of risks. The new approach incorporates an exponential, time-dependent component that generally reduces the risk margin for long-term liabilities, thereby moderating the capital intensity of long-duration contracts and bringing the prudential valuation closer to underlying economic reality. The cost-of-capital rate used in the calculation is set at 4.75% from 2027 onwards, with periodic reviews to ensure that it remains within a corridor between 4% and 5%. With respect to the best estimate of technical provisions, SNCU and undertakings that obtain prior supervisory approval may use a prudent deterministic valuation for certain life obligations with options and guarantees that are not considered material, enabling targeted simplifications without undermining reliability.

Within the SCR standard formula, several sub-modules are recalibrated. In the interest-rate risk sub-module, methodologies and parameters must reflect the possibility of further declines in interest rates

even when they are already low or negative, while excluding the risk of rates falling below a predetermined negative floor. The implementation of the revised interest-rate shock is phased in over a transitional period of up to five years to avoid abrupt changes in capital requirements. To incentivise genuinely long-term equity investments, the Directive facilitates the use of a more favourable standard parameter corresponding to an instantaneous loss of 22% for qualifying equity exposures held with an average holding period of more than five years, subject to stringent conditions demonstrating the undertaking's ability to avoid forced disposals in stressed market conditions. In addition, the symmetric adjustment to the equity capital charge is redesigned so that it cannot reduce or increase the equity requirement by more than 13 percentage points relative to its standard level, thereby dampening procyclicality. Undertakings may also rely on simplified calculations for risk modules or sub-modules deemed non-material, provided that each such module represents less than 5% of the Basic SCR and that the aggregate effect of simplifications does not exceed 10% of the Basic SCR, with a maximum duration of three years.

New macroprudential and liquidity-related instruments complement these quantitative adjustments. Undertakings must establish and regularly update a liquidity risk management plan covering the analysis of short-term cash flows and, where required by the supervisory authority, medium- and long-term flows as well. SNCU are exempt from the requirement to prepare such a plan. Where an undertaking's solvency position deteriorates, supervisory authorities may take preventive measures, including the request to update preventive recovery plans or to implement the measures set out in those plans. They may also require the suspension or restriction of variable remuneration, bonuses or capital distributions when these would jeopardise the undertaking's financial position. In exceptional circum-

stances that give rise to material liquidity risk, and only as a measure of last resort, supervisory authorities may temporarily suspend policyholders' surrender rights in life insurance contracts for up to three months, provided that the suspension serves the collective interest of policyholders and beneficiaries. During periods of exceptional sector-wide shocks, authorities may also require particularly vulnerable undertakings to limit or suspend distributions such as dividends, bonuses and share buy-backs.

The Directive reshapes the reporting and disclosure framework (Pillar III). The Regular Supervisory Report is generally submitted every three years, with the possibility of longer intervals for SNCU, and may also be prepared on a consolidated basis for groups. Submission deadlines are extended to 18 weeks for solo undertakings and 24 weeks for groups to allow for improved reporting quality. The Solvency and Financial Condition Report is restructured into two parts with distinct audiences. One part is addressed to policyholders and beneficiaries and provides essential information on the nature of the business, the main risks, including sustainability-related risks, and the capital position. The other part targets professional users and contains detailed technical data, including the calculation of SCR and MCR, the effects of long-term guarantee and transitional measures and exposures to climate-related risks. The audit requirement applies to undertakings that are neither SNCU nor captives, and flexibility is introduced with regard to reporting deadlines and procedures in exceptional circumstances.

At group level, supervision and consolidated solvency assessment are modernised and harmonised. The definition of a group is expanded to capture de facto groups and certain types of third-country holding structures, addressing configurations that may pose systemic or contagion risks but would otherwise escape comprehensive oversight. Group supervisors are granted powers to request structural modifications where the

organisation of the group impairs effective supervision. The notion of a small and non-complex group is introduced, enabling proportionality measures similar to those available to SNCU at solo level. The calculation of group solvency is clarified for participations in other financial sectors, which must be assessed in line with the requirements applicable to those sectors, and for immaterial participations and the application of the consolidated minimum capital requirement, including transitional arrangements. Macroprudential tools are extended to the group level, with requirements for group-wide liquidity management plans, the possibility to restrict intragroup transactions in exceptional situations and enhanced reporting obligations through consolidated supervisory and public reports, subject to audit. For groups whose ultimate parent is located in a non-equivalent third country, the group supervisor may require the establishment of an intermediate EU holding company or impose limits on transactions that could threaten the financial position of EU-authorised undertakings.

The updates introduced by Directive (EU) 2025/2 regarding the Internal Model, whether full or partial, are aimed at increasing harmonisation across undertakings' modelling approaches, enhancing the comparability of outcomes, and supporting a more effective integration of macroprudential considerations into the broader risk-assessment framework.

A key change compared with the regime established under Directive 2009 is the introduction of a mandatory requirement for undertakings using an Internal Model to periodically provide an estimate of the Solvency Capital Requirement calculated using the Standard Formula. Under the previous framework, such estimates could be requested at the discretion of the supervisory authority and subject to appropriate justification. The 2025 update makes this a systematic exercise for all Internal Model users, requiring them to submit, every two years, an SCR calculated in accordance with the

Standard Formula. The objective is to generate comparable indicators across undertakings using different modelling approaches and to facilitate longitudinal assessments of their evolving risk profiles, ensuring consistency with the methods and assumptions underlying the Standard Formula.

Directive (EU) 2025/2 also introduces stricter rules for the treatment within Internal Models of the impact of credit-spread movements on the Volatility Adjustment. When a supervisory authority authorises an undertaking to reflect such effects in its model, the resulting SCR may not fall below a predefined notional minimum. This floor must be calculated either using EIOPA's methodology or on the basis of the undertaking's asset portfolio. The intention is to prevent the VA from producing disproportionately large reductions in capital requirements and to preserve its role as a stabilising mechanism intended to mitigate excessive bond-spread volatility.

Furthermore, undertakings using an Internal Model must fully exclude the macro-component of the Volatility Adjustment. Including such a component within the Internal Model would materially weaken incentives to maintain resilience during systemic stress and would undermine the prudential purpose of the macro-component within the macro-prudential framework.

The changes introduced by Directive (EU) 2025/2 to the Standard Formula affect several core dimensions: the adjustments to the risk-free curve and to technical provisions, the calibration of SCR risk modules, and the reinforcement of the proportionality principle through explicit quantitative criteria.

- **Adjustments and technical provisions.**

- Risk Margin: the Risk Margin is subject to significant methodological and parametric revisions. From 30 January 2027, the cost-of-capital rate is set at 4.75% (currently 6%), with the

possibility of revision only from 31 January 2032 and within a 4 to 5% corridor. The calculation method is refined through the introduction of a time-dependent exponential factor, built on a time-adjusted SCR profile. This update provides a more realistic representation of the temporal decline of non-hedgeable risks and results in a lower Risk Margin, particularly for long-duration liabilities.

- Volatility Adjustment: the Volatility Adjustment framework is strengthened from both an authorisation and methodological perspective. Its application is now subject to prior supervisory approval. The calculation must include an undertaking-specific adjustment that incorporates the credit-spread sensitivity ratio of the asset portfolio and the interest-rate sensitivity of the best estimate of technical provisions. The proportion of the risk-corrected spread that can be added to the risk-free curve is increased to 85%. The previous country component is replaced with a macro-component defined by the difference between the euro-area risk-corrected spread and the country-specific risk-corrected spread, with the aim of reducing discontinuities and mitigating cliff-edge effects.
- Risk-free interest-rate term structure: the revised extrapolation methodology for the risk-free curve includes a phased-in transition mechanism to ensure a gradual implementation. This feature is particularly relevant for life insurers with long-term liabilities that are sensitive to changes in the extrapolated sec-

tion of the curve.

- **SCR risk modules.**

- Equity risk: the symmetric adjustment for equity risk is revised to allow a broader variation range, enabling adjustments of up to 13 percentage points relative to the standard charge, compared with the previous 10-point limit. The Directive also introduces a dedicated prudential treatment for long-term equity investments, allowing for a more favourable calibration under specific conditions. As a result, the duration-based equity sub-module under Article 304 of Directive 2009, already subject to restrictive criteria, is discontinued for new applications from 30 January 2027, with grandfathering for undertakings that were already applying it.
- Interest-rate risk: the interest-rate risk sub-module is recalibrated to reflect environments characterised by very low or negative interest rates. The revised calibration includes an explicit negative floor designed to avoid excessively large downward shocks to the curve and preserve realistic rate dynamics.

- **Proportionality and simplified calculation.**

- Simplifications for immaterial risks: Directive (EU) 2025/2 introduces explicit quantitative criteria defining when a risk module or sub-module may be considered immaterial. Undertakings may use simplified calculations if the module contributes no more than 5% to the basic SCR, subject to a cumulative ceiling of 10% and a maximum duration of three years. This ap-

proach allows undertakings to focus analytical efforts on material areas while maintaining robust capital assessment standards.

- Small and Non-Complex Undertakings (SNCU): the Directive formally introduces the category of small and non-complex undertakings, which automatically benefit from proportionality measures. These undertakings may use simplified calculations when a module or sub-module contributes less than 2% to the basic SCR, subject to the cumulative limit of 10%. This clarification supports a more consistent and coherent application of the proportionality principle across Member States.

Deep Dive Into Standard Approach and Internal Models in Insurance

With the entry into force of Solvency II, European insurance undertakings must hold enough capital to ensure that, with a probability of 99.5% over a one-year horizon, losses do not wipe out their own funds. The regulation provides two methodologies for calculating the capital requirement (SCR):

1. **Standard approach (Standard Formula):** a predefined method applicable to all firms using uniform regulatory parameters. It is based on risk modules (market, counterparty, life, non-life, health, operational, intangible) and uses fixed correlation coefficients. This is the default option for companies that do not have an approved internal model.
2. **Internal model (IM):** a quantitative model developed by the company to reflect its own risk profile. It requires

approval by the supervisory authority and must meet criteria for effective use, statistical quality, calibration, validation and documentation. The company is free to define risks, risk measures and aggregation methods [8].

The choice between the standard approach and an internal model influences the level of capital required, risk sensitivity and management strategies. The following sections present technical details of the two methods, with formulas, analogies and practical implications for a technically advanced audience.

The Standard Approach

The standard approach provides a module-based approximation of the Solvency Capital Requirement. It translates complex risk profiles into a modular structure supported by prescribed shocks and linear correlations.

At its core, the method offers a transparent and harmonised framework for capital calculation, enabling comparability across undertakings while relying on strong simplifying assumptions.

Modular Architecture and Variance-Covariance Method

The standard formula is built hierarchically. Each risk (e.g., market, counterparty, non-life) is decomposed into sub-modules (interest rate, equity, spread, catastrophe, etc.). For each sub-module an undiversified capital requirement SCR_i is calculated via volatility factors or shock scenarios. Aggregation is performed using the square-root formula:

$$SCR_{module} = \sqrt{\sum_{i,j} \rho_{ij} SCR_i SCR_j},$$

where ρ_{ij} is the correlation coefficient prescribed by the regulation between sub-modules i and j . For example, QIS5 for

life risk proposes correlations such as -0.25 between longevity and mortality risk or 0.5 between expenses and revision risk [2]. The same scheme applies at the higher level to aggregate modules into the Basic SCR and then add operational risk and the loss-absorbing capacity adjustment.

The key assumption of the previous formula is that losses are elliptically distributed (normal or lognormal) and dependence is linear. The square-root formula derives from an approximation of the total variance: if $L = \sum_i L_i$ with $\mathbb{E}[L_i] = 0$, then $Var(L) = \sum_i \rho_{i,j} \sigma_i \sigma_j$. Implicitly, the risk measure is assumed to be proportional to the standard deviation. Correlations include a prudence margin and are typically 1, 0.5 or 0.25 [3].

Non-life Risk Calculation: Combining Premium and Reserve

In the non-life module the standard formula estimates three components: premium risk, reserve risk and catastrophe risk. For each of the 12 lines of business (LOB) the standard deviations σ_{prem} and σ_{res} are determined from regulatory tables [3]. The two risks for each LOB are combined with a fixed correlation of 50%:

$$\sigma_{LOB} = \sqrt{\frac{Prem^2 \sigma_{prem}^2 + Res^2 \sigma_{res}^2 + Prem \cdot Res \cdot \sigma_{prem} \cdot \sigma_{res}}{(Prem + Res)^2}}.$$

A geographic diversification factor is then applied to reduce the volume if the portfolio is spread across multiple countries [3]. Aggregation across LOBs uses a 12x12 correlation matrix with entries 1, 0.5 or 0.25 [3]; the total standard deviation is:

$$\sigma_{tot} = \sqrt{\frac{1}{V_{tot}^2} \sum_{j,k} \rho_{jk} \sigma_j \sigma_k V_j V_k},$$

where V_j are the volumes (premiums + reserves) after diversification. Finally, the underwriting capital requirement is approximated as three times σ_{tot} , consistent with the 99.5% VaR of a lognormal distribu-

tion [3]. Analogous formulas exist for life, health and market modules, with predefined shock scenarios (e.g., $\pm 15\%$ on mortality rates, ± 100 bps interest-rate shock, 39% equity drop) [2]. Risk factors are calibrated at European level and represent a compromise between simplicity and prudence.

Conceptual Limitations and Issues

Although the standard formula offers a coherent and easily implementable framework, its underlying assumptions and structural design inevitably introduce some conceptual limitations. In practice, several recurring issues are well-known to practitioners:

- **Simplicity vs. realism:** the standard approach operates like a "pre-printed correlation matrix". All insurers use the same parameters regardless of portfolio composition. This ensures comparability, but penalises companies with highly diversified portfolios or strongly skewed distributions.
- **Normality and linearity:** the square-root formula and constant correlations assume losses follow an elliptical distribution and dependence is linear. In the presence of heavy tails or conditional correlations (tail dependency), aggregation may under- or over-estimate diversification [8].
- **Fixed modularity:** the structure of modules and sub-modules is predefined; risks cannot be refined further. The standard approach does not include expected profits or stochastic expenses and can be conservative for complex assets (e.g. equity-release mortgages, infrastructure).
- **Risk proxies:** the formula does not account for the full spectrum of emerging risks (cyber, climate change, ESG risks) and does not allow multi-year measurements. It is therefore suitable

as a rule of thumb but not as a strategic planning tool.

Internal Models

An internal model allows the SCR to be calculated using the company's own definitions of risks, measures and aggregation methods. According to EIOPA, a firm can:

- Define risk factors and build the full loss distribution instead of focusing solely on volatility [8];
- Choose risk measures other than VaR, such as Tail Value-at-Risk (TVaR) or Expected Shortfall, which are more sensitive to tails;
- Use longer time horizons to capture economic cycles;
- Model dependence via copulas (Gaussian, t-Student, vine copula) and introduce tail dependency, going beyond linearity [8];
- Adopt alternative aggregation schemes: bottom-up integration (joint simulation of all risks), modular approach (aggregating profits and losses at module level with calculated correlations) or sideways integration (aggregating cross-cutting risks) [8].

The flexibility of an internal model is comparable to designing a high-resolution navigation system: instead of following a pre-printed map, the firm builds a dynamic map based on its own data and chooses the desired level of detail. This freedom enables recognition of diversification between segments ignored by the standard formula and inclusion of mitigation strategies (reinsurance, derivatives, matching adjustment). Unlike the standard formula, internal models compute the risk measure directly from the simulated distribution. The 99.5% VaR is defined as $\inf\{x : F_L(x) \geq 0.995\}$, while Expected Shortfall (TVaR) is the conditional expectation of loss beyond the VaR: $TVaR_{99.5\%} = \mathbb{E}[L | L > VaR_{99.5\%}]$. Many

firms use TVaR because it penalises tails more and provides a more stable estimate. Risk aggregation can be performed via copulas; for example, a t-Student copula with 10 degrees of freedom captures greater tail dependency than the Gaussian. The internal model can also calibrate stochastic correlations that vary with market conditions. Sensitivity and capital contributions. The EIOPA study introduces risk multiplier and correlation multiplier indicators to assess sensitivity of diversified capital to individual risks and correlations:

- **Risk multiplier:** measures how much the diversified capital requirement changes in response to a small increase in an undiversified SCR. Formally, if SCR^{VC} is the diversified SCR and $undivSCR_i$ the undiversified requirement of risk i , the risk multiplier is approximately:

$$riskMultiplier_i = \sum_j \rho_{ij} \frac{undivSCR_i}{SCR^{VC}} \frac{undivSCR_j}{SCR^{VC}},$$

where $SCR^{VC} = \sqrt{\sum_i \sum_j \rho_{ij} undivSCR_i undivSCR_j}$ is the variance covariance diversified capital requirement. A risk multiplier of 50% for market risk implies that increasing its undiversified SCR by 2 units raises the diversified SCR by 1 unit [8].

- **Correlation multiplier:** evaluates how a small variation in the correlation ρ_{ij} between two risks affects the diversified capital. It is defined as the derivative of SCR_{div} with respect to ρ_{ij} , normalised by the diversified SCR [8]. High values indicate that dependence between those risks is a major driver; validating this requires sensitivity tests on alternative copula structures [17].

Advanced Modelling Techniques

For non-life risk, internal models use frequency-severity or collective risk models. Frequencies are modelled with Poisson, negative binomial or more complex count models; severities with lognormal, Pareto or Burr distributions. Monte Carlo simulation aggregates the number of claims and loss amounts to generate the loss distribution. For instance, an internal model may combine collective risk theory with log-normal bootstrap techniques for reserve risk, potentially achieving a more risk-sensitive capital assessment than the standard volatility-factor approach. Beyond simulation-based approaches, there is also an extensive actuarial and operational-risk literature on analytical and semi-analytical approximations for frequency-severity models. In the classical collective risk framework, the aggregate annual loss can be represented as a compound process:

$$S_T = \sum_{k=1}^{N_T} X_k,$$

where N_T denotes the number of loss events over the time horizon T , typically modelled as a Poisson or mixed Poisson process, and X_k are independent severity amounts. When $N_T \sim \text{Poisson}(\lambda T)$, the aggregate loss S_T follows a compound Poisson process. This representation is common in insurance risk modelling and has also been extensively developed in the banking literature on Operational Risk under the Basel framework, particularly through the Loss Distribution Approach, where frequency and severity distributions are estimated separately and then compounded to obtain the annual loss distribution.

The Operational Risk literature has provided important contributions to analytical approximations for high quantiles of aggregate loss distributions, especially in heavy-tailed settings where direct convolution is difficult and Monte Carlo simulation may require a very large number of scenarios to estimate extreme percentiles accurately.

Among these methods, the Single Loss Approximation provides closed-form asymptotic approximations of risk measures such as Value-at-Risk and Expected Shortfall by exploiting the fact that, for sub-exponential or heavy-tailed severity distributions, the tail of the aggregate loss is often dominated by the largest individual loss [18]. A first-order intuition is that, at high confidence levels, the aggregate loss quantile can be approximated through a transformed severity quantile:

$$VaR_{\alpha}(S_T) \approx F_X^{-1} \left(1 - \frac{1 - \alpha}{\lambda T} \right),$$

with refined versions adding correction terms depending on the expected number of losses and the severity moments. These results are particularly relevant when comparing numerical methods, such as Monte Carlo, Panjer recursion or Fourier-transform techniques, with analytical approximations for capital estimation. For market and balance sheet risk, internal models employ nested simulations: each real-world scenario (outer simulation) is linked to multiple risk-neutral scenarios to value assets and liabilities. This technique, known as Nested Monte Carlo, is computationally intensive. To reduce the burden, replicating portfolios and Least Squares Monte Carlo (LSMC) are used. LSMC approximates the value of assets and liabilities with regressions on a few inner simulations, producing accurate estimates in shorter times [16]. LSMC and replicating portfolios are now standard in life internal models to value products with options and guarantees.

Approval, Governance and Validation

To obtain approval of an internal model, a company must pass the Use Test (showing use in day-to-day management), meet statistical quality and calibration standards and document the model. The process includes:

- **Calibration:** deriving parameters from internal data (claims, exposures)

or external sources; for market risk, forward curves, implied volatilities and updated mortality tables are used. Supervisors require external benchmarks to avoid excessively optimistic calibrations [1].

- **Validation:** testing model adequacy bottom-up and top-down. Lloyd's 2025 guidelines emphasise that validating dependencies is complex and requires sensitivity tests on copula parameters and comparisons with historical data or stress tests [17]. It is necessary to justify the choice of dependency structures and show the presence (or absence) of tail dependency.
- **Profit & Loss attribution and model drift:** comparing simulated performances with actual results, analysing deviations and updating parameters. Documentation must be complete and available to the supervisor [17].

Practical Advantages and Disadvantages

Internal models offer clear advantages over the standard formula but introduce their own set of challenges. Their use therefore reflects a balance between greater accuracy and increased complexity. The main benefits and drawbacks typically recognised in practice are the following:

Advantages

- **Customisation and precision:** internal models capture portfolio-specific volatility, fully recognise diversification and include expected profits and stochastic expenses. Tail and copula modelling provides a more robust estimate of capital than the standard approach.
- **Capital optimisation:** numerous studies show that internal models reduce the SCR relative to the standard formula, with average reductions of 8.7%

for life companies. This benefits investments and asset-liability management strategies.

- **Support for risk management:** building the model enhances risk awareness, provides indicators for capital allocation and allows designing more efficient products and reinsurance [1]. The longer time horizon (>1 year) enables inclusion of climate and sustainability risks.

Disadvantages

- **Costs and complexity:** developing an internal model requires significant investment in systems, data and expertise. Even with proxy modelling techniques like LSMC, data management and maintenance remain onerous [1].
- **Regulatory oversight and conservatism:** supervisors may impose prudential margins or limit diversification benefits [1]. Approval is a formal process and the company is obliged to use the model until new approval is granted.
- **Modelling risk:** errors, wrong assumptions or unstable parameters can lead to misestimation of capital. Validation must address tail dependency and consistency with reality.
- **Comparability and transparency:** modelling freedom makes it difficult to compare results across firms. EIOPA has highlighted wide variation in diversification levels; metrics such as the risk multiplier are needed to understand which risks drive capital [8].

Analogies and Comparisons between the Two Approaches and Strategic Implications

Some practical and strategic implications emerge when comparing the two approaches, as the choice between standard

formula and internal model affects not only the level of capital but also pricing, risk management, investment decisions and the way results are communicated across the organisation:

- **Capital and pricing:** the standard approach tends to generate higher capital when loss distributions are asymmetric or when real correlations are lower than the prescribed ones. For companies with simple or monoline portfolios, this excess capital represents a safety margin and can reduce modelling burdens. However, for diversified firms or those with complex mitigation structures, extra capital may hurt competitiveness. Internal models allow pricing optimisation by accounting for capital costs per line of business and enabling targeted reinsurance strategies.
- **Risk management and asset allocation:** an internal model provides granular analysis for each risk driver, allowing identification of the most capital-consuming lines of business. Sensitivity metrics (risk multiplier) help understand which risks have the greatest impact on capital and evaluate the effectiveness of hedges [8]. In financial terms, replicating portfolios and LSMC techniques improve asset-liability matching, enabling optimisation of the investment portfolio composition and dynamic evaluation of the matching adjustment.
- **Reporting and communication:** the standard formula produces a result that is easily communicated to supervisors and shareholders. Internal models, while more accurate, generate complex outputs that require technical expertise to interpret. It is vital to create clear reports for management and ensure that governance understands key assumptions and sources of uncertainty.

Aspect	Standard Formula	Internal Model
Metaphor	A "cut-and-paste" method: the firm applies predefined parameters and correlations, akin to using a recipe book.	A custom lab: the firm builds its own model like an engineer designing an aircraft, choosing materials and testing interactions.
Definition of risks	Fixed modules and sub-modules with predefined volatility factors.	Risks defined by the firm; can include emerging risks, climate risks, options and guarantees [10].
Dependence	Constant linear correlations (0, 0.25, 0.5, 1). Do not capture tail dependency [5].	Flexible dependence structures: Gaussian or t-Student copulas, stochastic correlations, dynamic market factors. Can model tail dependence [10].
Risk measure	Based on standard deviation and lognormal approximation. Implicitly uses VaR at 99.5% via constant 3 [5].	Can adopt VaR, TVaR or other coherent measures. Uses the simulated distribution of results.
Diversification	Predefined by the correlation matrix; does not recognise complex diversification across new risks.	Calculated based on portfolio profile. Can fully recognise reinsurance and hedging strategies.
Expected profits and expenses	Not considered, leading to higher capital.	Included in the economic valuation.
Regulation and costs	No approval required; quick and low-cost implementation.	Requires approval and high costs; mandatory use once approved.
Governance	No need for complex internal structures; compliance is limited.	Requires use tests, validation, documentation, robust governance and ongoing review [18].

TABLE 1: Comparison of Standard Formula and Internal Model.

• **Key elements to be taken into account when choosing the approach:**

1. Size and complexity of the firm: large companies with multi-line portfolios and sophisticated strategies benefit more from internal models; smaller companies may prefer the standard approach to minimise costs;
2. Data availability and resources: an internal model requires long, granular time series, IT infrastructure and quantitative teams; without these resources, the standard formula remains the only option;
3. Capital benefit vs. costs: the company must assess whether the reduction in SCR justifies development costs and approval risk. The presence of expected profits not recognised by the

standard formula can be a decisive driver;

4. Mitigation strategies: heavy use of reinsurance, derivatives or asset-liability management techniques can only be fully valued through an internal model;
5. Regulatory outlook: more flexible regulations (like Solvency UK) simplify internal model approval [5]; firms should anticipate such developments.

Solvency II Review and Implications

At the end of 2023 European institutions reached a compromise on the review of Solvency II; the amended directive entered into force in January 2025 (see Section “Solvency II: Framework Evolution”) and member states have two years to transpose it, with full effect expected by 2027 [20]. This review aims to free up capital and encourage long-term

investment while maintaining policyholder protection. The main changes affect long-term measures and the interest-rate framework and include:

1. **Risk Margin reduction and introduction of λ :** the review lowers the Cost of Capital (CoC) from 6% to 4.75% and introduces a tapering factor (λ) to reduce the weight of long-term exposures. In the new scheme the Risk Margin is calculated by projecting the SCR for each year t , multiplying each projection by λ_t (with $0 \leq \lambda_t \leq 1$) and discounting at the risk-free rate [19]. EIOPA proposed an exponential curve with $\lambda_t = 0.975^t$ and a floor at 50%; the final compromise removed the floor but keeps the flexibility to define λ in Level 2 texts [15]. In a case study, applying a CoC of 4.75% and $\lambda=0.975$ reduces the Risk Margin by about 34%, freeing capital and boosting the solvency ratio by more than 20 percentage points [19]. Aegon notes that the exponential taper reduces interest-rate sensitivity of the Risk Margin by 40% for maturities of 20 years and brings the European CoC close to the UK regime, though still higher than in other jurisdictions [20].
2. **Volatility Adjustment (VA) revision:** the VA, added to the discount rate to dampen the impact of spread movements, is recalibrated. The general application ratio for the risk-corrected spread increases from 65% to 85%, and a Credit Spread Sensitivity Ratio (CSSR) is introduced to measure the alignment of the company's portfolio with the representative portfolio [19]. The new currency component formula becomes $VA_{\text{currency}} = 85\% \times CSSR_{\text{currency}} \times RCS_{\text{currency}}$ [20]. The volatility adjustment for the euro can be increased by a macro volatility adjustment, that depends on the country. A firm-specific adjustment is also

allowed, subject to supervisory approval, to use the firm's own portfolio spreads for two consecutive quarters, capped at 105% [15]. The case study company's simulations show that increasing the VA by 10 bps raises the solvency ratio by about 4 points by reducing the best estimate of reserves [19].

3. **Relaxing Long-Term Equity (LTE) constraints:** the LTE module provides a reduced capital charge (22%) for equities intended to be held long term. The review removes the ring-fencing requirement of assets and liabilities and reduces the minimum holding period from 12 to 5 years; it also extends eligibility to all OECD countries (not just EU/EEA) and requires adequate portfolio diversification [19]. Aegon reports that the new directive eliminates the need to couple the equity portfolio to a specific liability portfolio and limits the no forced-selling condition to five years [20]. In the case study mentioned above, reallocating 30% of equities into LTE reduces the market capital charge by about CHF 101 k and raises the solvency ratio by 5 points [19].
4. **Revising interest-rate risk and discount-curve extrapolation:** a new extrapolation method will replace the Smith-Wilson model: starting at the First Smoothing Point (FSP, 20 years for the euro), forward rates will converge gradually to the Ultimate Forward Rate (UFR) with a minimum weight of 77.5% after 40 years [15]. The proposed curve is slightly lower at long maturities and alters rate sensitivity; the transition will be gradual until 2032 [15]. In parallel, interest-rate shock parameters in the standard module have been revised: in addition to multiplicative factors, additive shocks are introduced and the 1% floor for up-rate scenarios is removed

[15]. A five-year transition period is envisaged for these new shocks.

5. **Width of the Symmetric Adjustment and other novelties:** the band for the symmetric adjustment on equity risk widens from $\pm 10\%$ to $\pm 13\%$, allowing a stronger counter-cyclical margin during extreme markets [20]; parameters for infrastructure assets and matching adjustment are revised to recognise diversification; criteria are introduced to classify small and non-complex undertakings (SNCU) with proportional reporting measures and reduced requirements [15]. In Pillar 2, the review introduces requirements on liquidity planning, diversity objectives and sustainability risks, while Pillar 3 provides more flexible reporting deadlines and certification by registered actuaries [15].

Implications for Internal Models and Standard Approach

The changes will have different impacts on the two approaches:

- **Standard formula:** the reduction of the Risk Margin and the higher VA will result in a lower solvency capital requirement, partly narrowing the gap with internal models. The expanded LTE eligibility and the new symmetric adjustment provide more flexibility in investment strategies. However, the complexity of the new VA (CSSR, macro-VA and firm-specific factor) may require more elaborate calculations and portfolio data collection, bringing the standard approach closer to a quasi-model approach.
- **Internal models:** the new framework allows firms with internal models to use dynamic VAs based on their own spreads and to model the curve extrapolation and Risk Margin with the

new λ . Copula and simulation techniques can incorporate the CSSR and the new interest-rate shocks to reflect the risk profile faithfully. However, the overall reduction of capital under the standard formula could reduce the competitive advantage of internal models; to justify their costs, firms will need to exploit the flexibility to include emerging risks, ESG planning and advanced reinsurance structures. The introduction of sustainability and liquidity measures in Pillar 2 aligns well with internal models, which can model multi-period scenarios and measure capital resilience.

In summary, the review of Solvency II is a step towards a more calibrated and sustainable regulatory framework, but its effectiveness will depend on the final calibration of parameters and the adoption of proportionate operational tools. Companies must carefully evaluate how to integrate the new rules into their capital management strategy and whether to retain or develop an internal model to maintain a competitive edge.

Industry Criticism and Comments

Despite the expected benefits, several associations believe the Level 2 proposals are not ambitious enough. Insurance Europe warned in 2025 that mis-calibration of the VA (especially the risk-correction component) and the new long-term interest-rate model could leave insurers exposed to artificial volatility, and that the review still limits the use of the LTE module. It emphasised that, while welcoming the reduction of the Risk Margin, Europe remains less competitive than other jurisdictions and the new rules introduce additional operational complexity and administrative burdens. The association therefore calls for further refinement of calibrations to generate growth and free up capital.

Further Developments

- **Regulatory reforms and proportionality:** the review of Solvency II and the adoption of Solvency UK will introduce greater proportionality and simplification. The shift from an "approval" regime to a "permission" regime reduces documentation burdens and accelerates authorisation, encouraging internal model adoption [5]. In parallel, EIOPA is studying metrics to compare diversification levels among models and may issue guidelines for comparability [8].
- **Integration of emerging risks:** Supervisors will require integration of climate, ESG and cyber risks in both the standard formula and internal models. The standard approach may be updated with new factors or scenario shocks; internal models, thanks to their flexibility, can already incorporate climate transition scenarios or sustainability-related shocks. Convergence between Solvency II and IFRS 17 means valuation models must align with accounting; internal models facilitate this integration.
- **Technological innovation:** the use of machine learning and artificial intelligence (e.g., neural networks for LSMC, deep learning models for replicating portfolios) is improving the ability to model losses and reduce computation times. Research is underway on transformer-based and explainable models for risk quantification. Authorities are exploring digital tools to monitor model adequacy in real time and facilitate continuous validation. 

References

- [1] **Chaplin, R. A.** *The Standard Formula: A Guide to Solvency II - Chapter 9: Internal Models*. Skadden Publication, July 2024.
- [2] **Christiansen, M.C.** *A sensitivity analysis concept for life insurance with respect to a valuation basis of infinite dimension*. Insurance: Mathematics and Economics, April 2008.
- [3] **Cofield, J., Kaufman, A. and Zhou, C.** *Solvency II Standard Formula and NAIC Risk-Based Capital (RBC)*. Casualty Actuarial Society, Fall 2012.
- [4] **Consiglio dei Ministri.** *Decreto Legislativo 3 novembre 2003, n. 307*. Gazzetta Ufficiale della Repubblica Italiana, November 2003.
- [5] **Dooley, A.** *Internal Models Under Solvency UK: Challenges and Opportunities*. Insurance Insights, July 2024.
- [6] **EIOPA.** *Insurance Risk Dashboard*. EIOPA-BoS-25-174, April 2025.
- [7] **EIOPA.** *Peer Review Work Plan 2023-2024*. EIOPA-BoS-22/345, June 2022.
- [8] **EIOPA.** *Study on Diversification in Internal Models*. January 2024.
- [9] **EIOPA.** *Technical advice on the implementation of the new proportionality framework under Solvency II*. EIOPA-BoS-24-590, January 2025.
- [10] **European Parliament and the Council.** *Directive 2002/12/EC*. Official Journal of the European Communities, March 2002.
- [11] **European Parliament and the Council.** *Directive 2009/138/EC*. Official Journal of the European Union, November 2009.
- [12] **European Parliament and the Council.** *Directive 2014/51/EU*. Official Journal of the European Union, April 2014.
- [13] **European Parliament and the Council.** *Commission Delegated Regulation (EU) 2015/35*. Official Journal of the European Union, October 2014.
- [14] **European Parliament and the Council.** *Directive (EU) 2025/2*. Official Journal of the European Union, November 2024.
- [15] **Garg, D. and Burke, S.** *Solvency II Amendments*. Finalyse, June 2024.
- [16] **Leitschkis, M.** *Solvency II proxy modelling via Least Squares Monte Carlo*. Milliman, January 2012.
- [17] **Lloyd.** *Internal Model Validation Guidance*. Lloyd's Guidance on Solvency UK Internal Model Validation, April 2025.
- [18] **Peters, G. W. and Shevchenko, P. V.** *Single Loss Closed-Form Approximations of Risk Measures*. In *Advances in Heavy Tailed Risk Modeling: A Handbook of Operational Risk*, Chapter 8, pp. 433–516. John Wiley & Sons, 2015.
- [19] **PwC.** *Unlocking opportunities under Solvency II Revisions effective 2027.*, 2025.
- [20] **Van Bragt, D.** *The Solvency II review has finished - What will change?* AEGON Insights, September 2024.



Innovation

**Understanding and Governing
Agentic AI: Foundations, Risks
and Control Frameworks**

About the Authors



Antonio Menegon:

Chief Innovation Office

He holds bachelor and master degrees in mathematics, with a specialization in quantitative finance, complemented by post-degree studies in artificial intelligence and machine learning. He began his career as a risk quant and risk analyst before moving into leadership role advising major banking groups. He guided teams across risk management, front office, and ICT departments, driving model development, risk assessments and monitoring, digital transformation, governance, and AI initiatives. He leads the Group's Innovation Center, focusing on the application of technology and artificial intelligence in financial services.



Valerio Ciminelli:

Business Analyst

He holds a MSc in Finance, Intermediaries and Markets from Alma Mater Studiorum University of Bologna. He has experience in the management of complex IT and strategic projects for major clients in the Financial sector gained working closely with Securities Services and Credit Management.





Margherita Ranieri:

Project Manager

She holds an MSc in Management Engineering with a specialization in Finance from the Politecnico di Milano. She is currently working as a Project Manager, actively engaged in Business and IT process implementation projects at one of Europe's leading banking institutions.



Agentic AI marks a shift from systems that generate content to systems that pursue goals, being able to plan, reason, and act with a degree of autonomy. In this paper, Generative AI primarily refers to the capability to produce new content, whereas Agentic AI refers to the system-level use of generative or reasoning capabilities within a goal-directed decision-and-action loop; the two are therefore distinct but not mutually exclusive, consistently with the terminology adopted by the Financial Stability Board. For financial institutions this shift is not merely technological: every increment of autonomy is also an increment of risk, and the controls designed for classical AI models are no longer enough. This paper sets out the risk landscape and governance requirements, while highlighting guidelines for deploying AI agents in regulated financial contexts. It makes two connected arguments. First, that control over the model and the stack, favoring open-weight models specialized through fine-tuning and deployed under the institution's control, is a suggested precondition for governable agentic AI, not only architectural and IT preference. Second, emerging supervisory and regulatory frameworks collectively confirm the need for a dedicated governance approach for agentic AI. While the April 2026 U.S. interagency guidance (SR 26-2), updating the traditional SR 11-7 Model Risk Management framework, explicitly excludes generative and agentic AI from its scope and directs institutions to govern such systems through broader risk-management practices, other frameworks, including the EU AI Act, DORA, and the UK PRA SS1/23, similarly strengthen expectations around lifecycle risk management, operational resilience, third-party governance, continuous monitoring, and human oversight. Together, these developments highlight a governance gap that existing model risk management frameworks alone cannot fully address. That gap is the mandate for a dedicated governance layer. This paper proposes the **Agent Risk Management (ARM)** framework as a governance and control system that evaluates not only outputs but the entire decision-making trajectory of autonomous systems, organized around four risk dimensions (Data, Vendor, Architecture, Human Factor) and operated as a continuous, evaluation-driven lifecycle rather than a point-in-time validation. This paper sets out the problem and the general structure of the proposed solution; the detailed methodology through which ARM is calibrated and operated is presented and tailored through direct engagement with each institution [4, 5, 14, 9, 8].

AGENTIC AI is among the more consequential turning points in enterprise automation. Unlike systems that execute predefined logic or respond reactively to prompts, agentic systems perceive their environment, reason over possible courses of action, and act, with limited supervision, to pursue user- or system-level objectives. What distinguishes them is not the individual task but the chaining: an agent breaks a goal into steps, calls tools and systems in sequence, observes the results, and sustains a task over many actions, even for hours, rather than automating a single one. For a financial institution, the question is no longer whether the technology is accessible, but whether it is governable. The properties that make agents valuable such as autonomy, tool use, persis-

tence, the ability to chain actions are precisely what make them hard to govern with existing controls. A system that can act on the world can also act incorrectly on it, and the consequences are no longer confined to a flawed output: they propagate into operational processes, security posture, and regulatory compliance. This is not a speculative concern, and timing is what makes it pressing. Agentic systems are moving beyond pilots into day-to-day operations drafting documents and memos, supporting daily and routine analysis, assisting in software development, etc. Their reach is widening toward the processes through which banks serve customers and manage risk, yet the supervisory apparatus that governs models in these institutions, and the foundational Model Risk Management (MRM)

guidance in particular, was not built for systems that decide and act (Section “*From Model Risk Management to Agent Risk Management*”). The result is a widening distance between what the technology can do and what the control framework was designed to oversee. This paper is concerned with that distance, not to advocate for a slowing adoption, but by setting out the understanding and the controls that let autonomy be extended deliberately and governed in an auditable and accountable manner: the foundations (Section “*Foundations of Agentic AI*”), the architectural choices that are themselves risk choices (Section “*Implementation as a Risk Decision: The Case for a Controllable Stack*”), the risk landscape and its regulatory anchoring (Section “*The Emergent Risk Landscape and Its Regulatory Anchoring*”), and the governance shift from Model Risk Management to Agent Risk Management (Sections “*From Model Risk Management to Agent Risk Management*” and “*The Pillars of an ARM*”).

Foundations of Agentic AI

This section assumes familiarity with the main concepts around large language models and the basic notion of an agent, hence we keep this section deliberately brief: its purpose is to fix the vocabulary and the structural picture that the risk analysis later depends on.

From Generative AI to Agentic Systems

It has become common to equate Generative AI with LLMs, but the term covers a wider spectrum of AI models; e.g. Generative Adversarial Networks (GANs), Variational Auto Encoders (VAEs), and diffusion models generate from latent or numerical inputs, not prompts. **LLMs** are just one member of that family, yet the one at the heart of agentic systems: they produce text or multimodal outputs in response to a query but remain reactive, responding when queried, without persistence, goal ori-

entation, or the ability to act. In this sense, Generative AI and Agentic AI should not be understood as mutually exclusive model families. Generative AI primarily denotes a capability, the production of new content, whereas Agentic AI denotes a system-level configuration in which generative or reasoning capabilities are embedded in a goal-directed process that can plan, select tools, and take actions with limited human oversight. This working distinction is consistent with the Financial Stability Board’s 2026 terminology[8], which treats Agentic AI as distinct from GenAI while recognising that agentic systems may incorporate generative capabilities. From this reactive core, systems are distinguished by two things: how much control is delegated to the model (autonomy) and how many agents coordinate (scale). Along these axes, three architectural forms of rising autonomy and complexity emerge, each conflated in practice, yet each introducing a distinct tier of risk:

- **Agentic Workflow (orchestrated autonomy):** an LLM (or a set of tools and models) executes iterative steps such as reflection, self-correction, or planning, but within predefined programmatic paths. The control flow is owned by human-written software logic; the model fills the steps but does not decide them. Autonomy is bounded by design: the defining feature is that the code, not the model, holds the reins (control–flexibility trade-off).
- **AI Agent (self-directed autonomy):** a single system in which the foundation model itself owns the control flow, interpreting a goal and deciding which steps to take and which tools to call to achieve it, wrapped in an infrastructure of tool-calling APIs, memory, and planning loops. Control passes from fixed code to the model, the step that turns “guided” into “autonomous.”
- **Multi-Agent System (MAS) (systemic autonomy):** an An intercon-

nected network in which multiple specialized, heterogeneous agents interact, collaborate, and negotiate over shared environments or messaging layers to pursue distributed goals. Intelligence and behaviour shift from the individual entity to ecosystem-level dynamics, introducing emergent risks that cannot be inferred from any single component.

Recent surveys observe a similar architectural evolution, where increasing levels of planning, memory, tool integration and autonomous execution progressively distinguish agentic systems from standalone generative models.[1] In summary, the reactive LLM provides the reasoning core; agentic workflows put it to work along fixed, orchestrated paths; AI agents hand the control flow to the model itself for self-directed execution; and multi-agent systems scale this into coordinated ecosystems. This progression, from content generation and reasoning, to guided execution, to autonomous action, and ultimately to distributed coordination, corresponds to a parallel escalation in the operational impact and governance requirements of AI-related risks (Section “*The Emergent Risk Landscape and Its Regulatory Anchoring*”).

Agents in a Nutshell

Agents can be classified along a continuum of autonomy (from simple reflex agents to fully autonomous agents; Russell & Norvig, 2021) [15], and their reasoning can be elicited and structured in various ways (Chain-of-Thought, Tree-of-Thought) or built into dedicated reasoning models. Regardless of the specific classification, a modern agent is not a monolithic model but a structured system, usefully described along two dimensions. From a governance perspective, agentic status therefore attaches to the deployed system rather than to the foundation model alone: the same generative model may support a non-agentic application or, when combined with

planning, memory, tool use, and delegated action, an agentic one, consistently with the distinction adopted by the Financial Stability Board [8].

The Core - The Static Resources the Agent Draws upon:

- **Instructions:** the explicit action plan governing how the agent approaches a task (shaped through prompting and, where needed, fine-tuning).
- **Knowledge Base and Memory:** short-term (working) memory, typically the context window, and long-term memory, typically vector stores enabling retrieval across sessions.
- **AI Model(s):** the foundation model that serves as reasoning engine. Two of its properties carry direct governance weight. First, it is **intrinsically stochastic**: outputs are sampled, so the same input need not yield the same output, and behaviour is non-deterministic by construction. For a regulated institution this is the most consequential property of all, not because its impact is always the largest, but because it breaks the assumption on which classical validation rests: a stable, reproducible input-to-output mapping that can be tested once and relied upon thereafter. Second, reasoning built on next-token prediction remains prone to characteristic failures (e.g., confidently invalid steps, degradation over long horizons, brittleness to superficial reformulation) and its depth is not uniform, depending on the class of model and on instruction design. The choice and control of the model therefore carry consequences that surface throughout the risk discussion.
- **Tools:** external APIs, models, scripts and functions through which the agent acts. Tools extend capability and reinforce auditability (sensi-

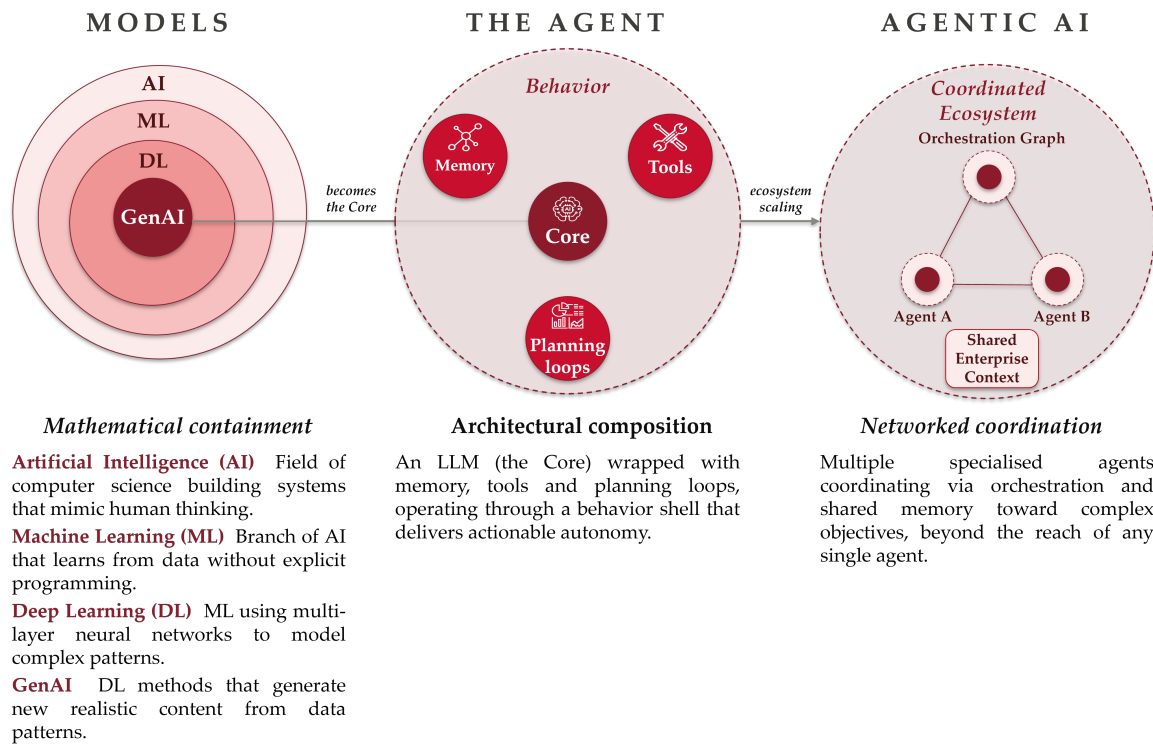


FIGURE 2: The Evolution of Generative Ai into Agenic Systems.

tive operations delegated to validated components are more predictable and traceable), but the same reliance expands the risk surface.

The Behaviour - Dynamic Loop through Which Those Resources Are Mobilised:

- The agent **Processes** raw inputs into structured understanding;
- **Plans** a strategy by decomposing the objective;
- and **Executes** using its model and tools, monitoring results and revising as needed.

This is the conceptual counterpart of the **Re-Act** paradigm (Reasoning + Acting), which interleaves reasoning, tool invocations, and observations in a single loop, letting the agent build and revise a plan while drawing on the environment for information (Yao et al., 2023) [16]. Left to reason unchecked, a model can produce a chain of fluent, confident steps that never touch reality and still arrive at a wrong answer. Grounding

each step in an actual tool result curbs this, the agent verifies against the environment as it goes, but does not eliminate it: the model can still misread a result or reason poorly about it. Beyond reasoning error, the loop also carries operational failure modes that later shape the risk picture: as context accumulates, token cost rises and model performance degrades, and the agent can fall into unproductive cycles or lose sight of its original objective. The “Core” and the “Behaviour” together explain what an agent is, how it acts, where its risks originate, and how, with rising autonomy, they propagate. Stochasticity and autonomy, not linguistic competence, are what most shape the risk profile (Section “From Model Risk Management to Agent Risk Management”).

Implementation as a Risk Decision: The Case for a Controllable Stack

Having characterised what an agent is, now we turn on how an agent is built since it

largely determines the extent to which, and the manner in which, it can be governed. The dominant mode of adoption visible today is to build on commercial, general-purpose models accessed through an API: an immediate route to capability and fast time-to-market. This is a genuine strength, and in a financial institution it has a natural place: rapid prototyping, moving from idea to a minimum viable product, and exploring where agentic automation adds value before committing to build. The question is what happens next. As such systems move from experimentation into processes that matter, the binding constraints become accuracy, predictability, cost control, data security, and integration into regulated core processes, and on each of these, the API-first model shows structural limits:

- **Security and vendor dependence.** API usage routes proprietary data through external infrastructure. Enterprise SLAs (e.g., data isolation, no-training guarantees, residency clauses) and institution-owned access layers (e.g., gateways enforcing guardrails, API controls, token governance) can materially reduce this exposure, but they mitigate rather than remove it: residency remains contractual rather than architectural, and single-provider reliance still creates lock-in whose terms, pricing, and model availability can change unilaterally.
- **Jurisdictional and continuity risk¹.** Beyond commercial decisions, model availability is increasingly subject to sovereign control. Governments can restrict, embargo, or block the cross-border distribution of advanced models through export controls and national-security measures, overriding any commercial agreement be-

tween institution and vendor. A model relied upon in production may therefore become unavailable not because the provider chose to retire it, but because a jurisdiction decided it could no longer be supplied, a contingency no SLA can insure against, and one that falls hardest on institutions whose critical processes are coupled to a model they neither host nor control.

- **Governance and reproducibility.** Commercial endpoints are often opaque and are periodically updated; an identical input need not yield an identical output over time, and a decision taken months earlier may be impossible to reproduce once the provider has revised or retired the model. Provider-managed aliases and automatic upgrades can silently change the model underlying an application. For processes subject to supervisory scrutiny, this undermines lineage, repeatability, and effective challenge.

The reframe. Industry attention mostly fixates on the capability race: which newly released frontier model now tops the benchmarks and leaderboards. For most production tasks the relevant question is different: which model is good enough for this task, at what cost, and under what degree of control. This matters because the capability gap between open-weight and closed-weight models has narrowed dramatically, from well over a year in 2023 to a matter of months by 2026 (Epoch AI tracking – see Figure 3). For a growing range of well-defined tasks, an open-weight model deployed under the institution's own control is no longer a compromise on capability. This trade off becomes evident when examining the current model landscape (Fig-

¹Anthropic's **Mythos 5** and **Fable 5** models, along with OpenAI's **GPT-5.6**, were abruptly blocked by the U.S. government in June 2026 due to national security concerns and export control directives. The White House cited risks of vulnerabilities that could allow AI models to be "jailbroken" to exploit software and cybersecurity codes.

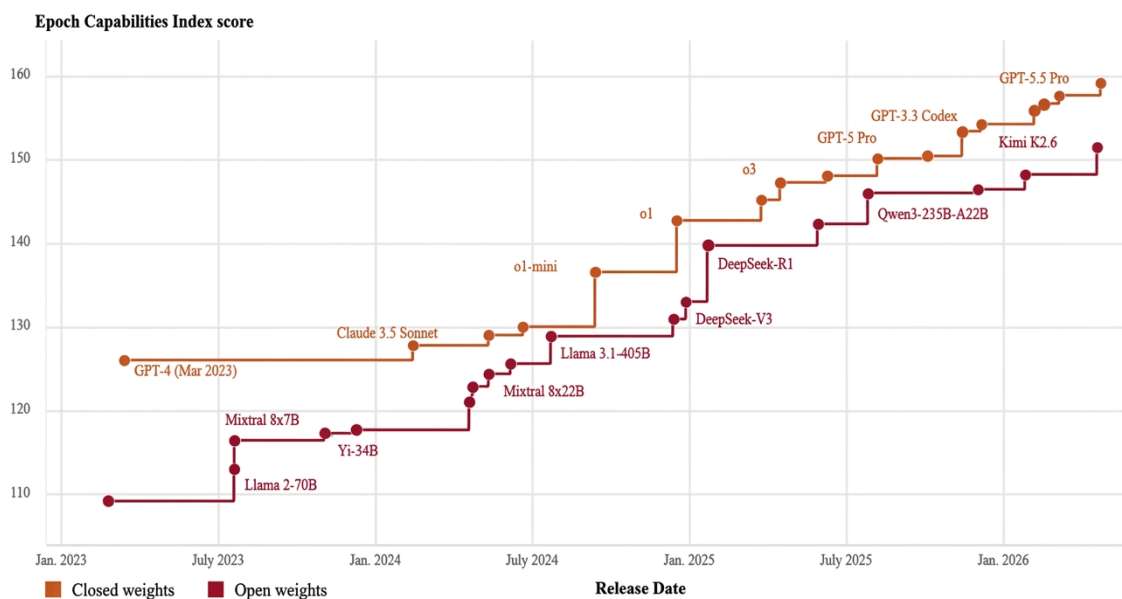


FIGURE 3: The narrowing capability gap between the best open-weight and frontier closed-weight models. Stylised representation based on Epoch AI's tracking of open- vs. closed-weight model performance [19].

ure 4): at any given level of capability, cost per task spans more than an order of magnitude, so frontier pricing buys little marginal capability for the vast majority of production tasks. Notably, the attractive region, high capability at low cost, is populated largely by open-weight models, while the absolute capability frontier remains both closed and expensive, outside the range most production tasks require. At the same time, the economic rationale for ownership is not automatic. The cost advantage of a self-hosted open-weight model depends on the efficient use of the underlying infrastructure, including GPU utilisation rates, inference volume, and the extent to which fine-tuning and operational costs can be amortised across that volume. Where utilisation is low or intermittent, a metered API may remain the more cost-effective option. Conversely, the ownership case becomes stronger as usage patterns become sustained, predictable, and sufficiently scaled. The key point is therefore not that open-weight models are intrinsically cheaper in all circumstances, but that they transform an opaque, per-call operating expense into a controllable capital and infrastructure decision that the institution

can actively govern and optimise.

The adoption of open-weight models, whose parameters are publicly available and can therefore be downloaded, inspected, fine-tuned, and deployed within the institution's own controlled environment, reintroduces a level of transparency, governance, and operational control that is structurally limited in API-first deployment models.

Under this configuration, **data remains within the institution's trust boundary, the model can be treated as a controllable technological asset rather than a metered external service, model versions can be frozen and governed through internal change-management processes, and operational continuity is less dependent on the commercial, technical, or jurisdictional decisions of a third-party provider.** Crucially, this control is independent of where the model runs: because the model is the institution's own, it can be deployed on-premise or on any cloud of choice (and moved between them) without surrendering ownership, reproducibility, or data control. For regulated institutions, especially those under on-premise or data-residency constraints, this portability of control is in-

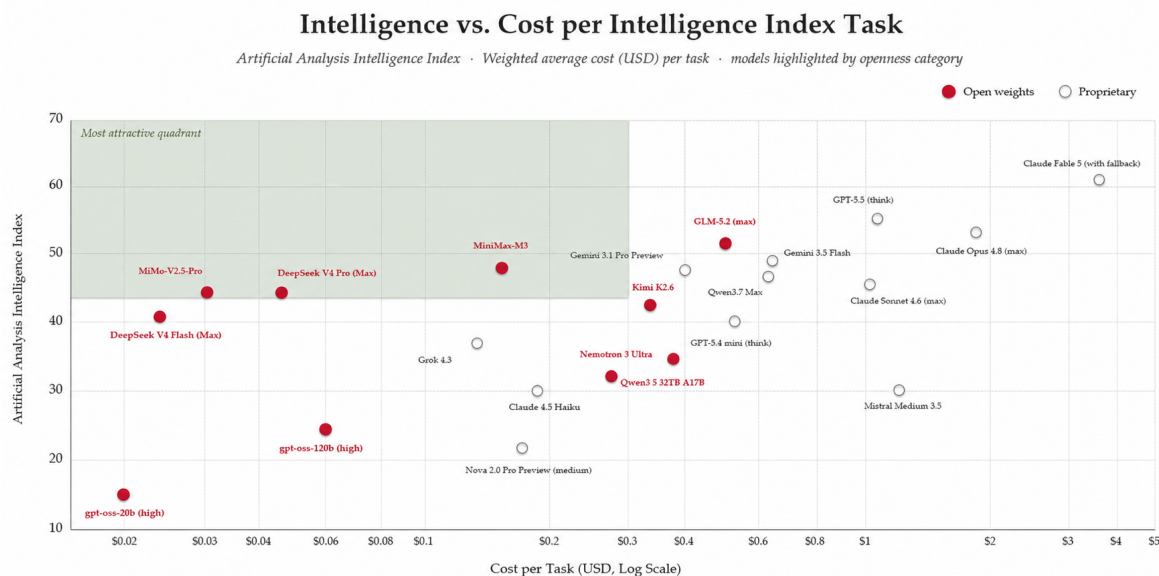


FIGURE 4: Intelligence versus cost per task across current models. Each point is a model; the horizontal axis is weighted average cost per task [18] (USD, log scale), the vertical axis the Artificial Analysis Intelligence Index² [18]. The spread at any given intelligence level illustrates that comparable capability is available across a wide cost range, and that the low-cost / high-capability region is dominated by open-weight models [17].

creasingly a precondition rather than an option.

This also reframes what "intelligence" means for an agent. A production agent does not need the highest possible score on general knowledge; it needs to interpret the request correctly and call the right tools with the right parameters [12]. These two capabilities, robust instruction-following and reliable tool-calling, are what allow routine actions to be delegated to validated components rather than re-derived in free-form generation, improving determinism and auditability. Fine-tuning a selected open-weight model on a curated, domain-specific dataset embeds the institution's language, taxonomies, and processes directly into the model, yielding an agent the institution genuinely owns and can govern over time.

For a regulated institution, control over the model and the stack is the foundation on which every downstream governance control depends. Where that control is absent,

governance is performed on a system the institution does not fully see and cannot reproduce.

The Emergent Risk Landscape and Its Regulatory Anchoring

This chapter reviews the principal sources of risk associated with advanced AI systems as identified in the literature and current supervisory frameworks, and organizes them into a **common analytical structure** that supports the development of the proposed Agent Risk Management (ARM) framework. Rather than introducing a new taxonomy of AI risks, the chapter synthesizes recurring themes emerging from both academic research and regulatory guidance into **four structural risk dimensions** and complements them with an operational perspective based on the stages of the agentic execution process. This synthesis provides the analytical foundation upon which the governance

²Including GDPval-AA v2, τ^3 -Banking, Terminal-Bench v2.1, SciCode, Humanity's Last Exam, GPQA Diamond, CritPt, AA-Omniscience, AA-LCR. Weighted average cost per task is calculated from input, cache hit, cache write, reasoning, and answer token prices, divided by task count, and weighted by its Intelligence Index weight. See Intelligence Index methodology for further details.

framework is subsequently developed.

Four Risk Dimensions

This work adopts four structural risk dimensions that remain applicable throughout the AI lifecycle, regardless of the underlying architectural complexity. These dimensions have been selected because they map closely with the principal risk domains emphasized by current regulatory and supervisory frameworks (see Section “*Regulatory Anchoring*”) and AI Governance principles and best-practice frameworks (such as the NIST AI Risk Management Framework [10] and the OECD AI Principles [11]):

- **Data Risks:** the quality, integrity, and governance of the information supporting the system.
- **Vendor Risks:** dependencies on external providers, foundation models, cloud infrastructure, and third-party services. For an institution relying on commercial models, this dimension absorbs much of what was traditionally “model risk.” Crucially, adopting open-weight models reduces this dependency but does not remove it: it converts a live dependency (metered access, silent updates, availability) into an inherited-provenance one: a fine-tuned model is an adaptation of a base the institution did not build and cannot fully audit (training data, latent capabilities, and safety properties are only partially observable). Vendor risk therefore shifts from the provider’s ongoing control toward the opacity of the inherited artifact, rather than disappearing.
- **Architecture Risks:** the design, orchestration, and operational behaviour of the system.
- **Human Factor Risks:** the interaction between humans and AI: oversight, accountability, and behavioral responses to automated decisions.

These dimensions should not be interpreted as mutually exclusive categories but as **complementary analytical perspectives** through which LLMs and Agent-related risks can be examined. Individual failure modes frequently span multiple dimensions simultaneously. For example, a prompt injection attack may originate from malicious input data, exploit vulnerabilities in the underlying model or external components, and ultimately compromise the execution logic of an agent interacting with external tools. Recognizing these interdependencies provides a more comprehensive understanding of AI risk and supports the design of governance mechanisms capable of addressing complex, cross-cutting vulnerabilities.

Operational Manifestation of AI Risks

While the four structural dimensions identify the primary sources of AI risk, they do not indicate where these risks emerge during the operation of an agentic system. To complement the proposed taxonomy, risks can also be analyzed from an operational perspective by mapping them onto the main stages of the agent execution process: information acquisition, reasoning and planning, and action execution. This dual perspective facilitates the identification of appropriate control mechanisms, as different risks become more prominent at different stages of the execution lifecycle. As autonomy increases, the primary focus of risk progressively shifts from the quality of information processing and reasoning toward the integrity and consequences of autonomous actions. Consequently, governance approaches centered exclusively on output quality become insufficient. Effective oversight must extend across the entire execution process, encompassing reasoning, planning, tool interaction, and action execution. Governance frameworks designed primarily to address risks arising during the early stages of the execution process, such as information retrieval and output

generation, are unlikely to provide adequate control over the operational risks associated with autonomous decision-making and real-world execution. This is especially true when systems scale from a standalone LLM to an agent to a multi-agent system. The risks remain similar, but their severity escalates rather than merely shifts (e.g., informational risk at the LLM level becomes operational at the agent level and systemic at the multi-agent level (Figure 5)).

Regulatory Anchoring

The four risk dimensions are consistent with the principal supervisory expectations emerging from current AI- and model-governance frameworks, which broadly converge around four governance domains: data quality and model reliability; third-party and vendor governance; operational robustness and resilience; and human oversight and accountability.

From the perspective of this research, the most significant regulatory observation is not simply that existing frameworks prescribe governance controls, but that the principal model-governance framework explicitly defines the boundary of its applicability. SR 26-2 [2], which replaced SR 11-7 as the primary U.S. supervisory framework for Model Risk Management, explicitly excludes generative and agentic AI from its scope³. The guidance specifies that its principles apply to "traditional statistical and quantitative models and non-generative, non-agentic AI models," while stating that broader institutional governance practices should determine the appropriate controls for systems falling outside its scope. Consequently, the principal supervisory framework for Model Risk Management explicitly recognizes that generative and agentic AI require governance mechanisms beyond the traditional MRM perimeter, leaving in-

stitutions responsible for defining appropriate controls within their broader enterprise risk-management frameworks. On the international stage, other foundational frameworks extend comparable governance expectations to AI and machine learning systems by emphasizing lifecycle risk management, operational resilience, human oversight, and third-party governance. However, their specific application to advanced configurations mandates a precise alignment with their respective implementation timelines and updated mandates:

- **EU AI Act [6]:** Establishes a highly structured, progressive enforcement timeline. Regulated entities must note that rules for General Purpose AI (GPAI) models became applicable on **August 2, 2025**, followed by the European Commission's direct enforcement capabilities starting **August 2, 2026**, with full progressive application continuing through **2027**.
- **Digital Operational Resilience Act (DORA) [7]:** Legally applicable since **January 17, 2025**, DORA's central intersection with agentic AI lies in its stringent requirements for **ICT third-party risk management and operational resilience**. The dynamic nature of autonomous agents makes them a key focus point under DORA's vendor compliance and supply-chain oversight protocols.
- **PRA SS1/23 [13]:** In the UK, the Prudential Regulation Authority's Supervisory Statement **SS1/23**, originally introduced in 2023, has been comprehensively updated, finalized, and published with revised supervisory enforcement protocols active in **2026**, reinforcing stricter institutional accountability and data controls for adaptive

³"Generative AI and agentic AI models are novel and rapidly evolving. As such, they are not within the scope of this guidance. Nonetheless, a banking organization's risk management and governance practices should guide the determination of appropriate governance and controls for any tools, processes, or systems not covered in this document. However, the principles described in this guidance apply to traditional statistical and quantitative models and non-generative, non-agentic AI models."

Dimension	Process (perceive / retrieve)	Plan (reason / decompose)	Execute (tool use / act)
Data	Incorrect, incomplete, outdated or poisoned information retrieved from internal or external sources.	Decision based on incomplete or biased context.	Ungrounded or incorrect information propagated into operational actions.
Vendor	Changes in external models or retrieval services that compromise semantic interpretation, output consistency, and system-level model availability.	Reasoning depth contingent on opaque, updatable model.	API latency / failures; or silent model updates affecting execution.
Architecture	Context management failures; memory inconsistencies and perception errors.	Goal misalignment; flawed or looping plans.	Prompt injection causing unauthorized tool calls; execution errors; cascading failures.
Human Factor	Over-trust in retrieved context.	Automation bias and insufficient human challenge of agent reasoning.	Excessive delegation of consequential decisions and actions; accountability gaps.

TABLE 2: Manifestation of the Four Risk Dimensions Across the Agentic Operation Process.

modeling structures.

Because none of these individual frameworks provide a cohesive, off-the-shelf methodology tailored specifically for autonomous agentic systems, the regulatory carve-out formalized by SR 26-2 **motivates the introduction of a dedicated Agent Risk Management (ARM)** framework capable of operationalizing governance requirements for increasingly autonomous AI architectures.

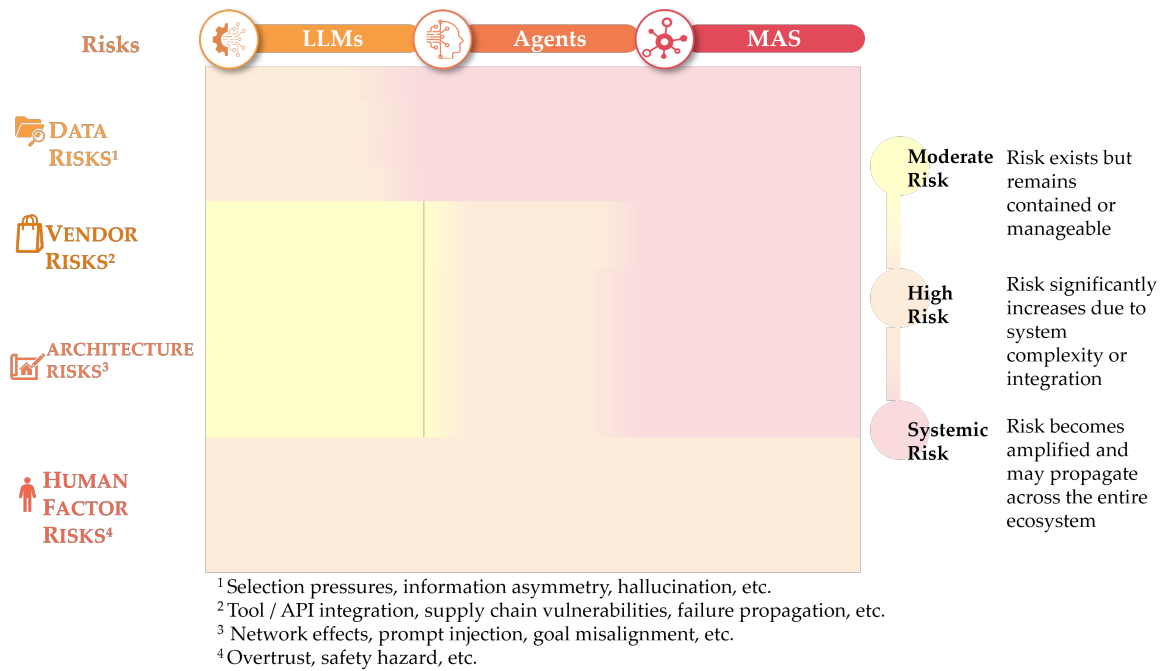


FIGURE 5: Heatmap of Risk Amplification across AI Architecture.

From Model Risk Management to Agent Risk Management

The Need for a Paradigm Shift

Historically, AI-related risk has been addressed through **Model Risk Management (MRM)** and its extension with Machine Learning (ML) Validation frameworks [3]. Three characteristics of agentic systems suggest the need for a further evolution of existing governance approaches:

1. External dependence: institutions increasingly rely on externally managed foundation models, reducing visibility into internals and changing processes;
2. Continuous interaction: agents interact with users, data, tools, and environments continuously, making behaviour difficult to assess through periodic (i.e., point in time) validation alone;
3. Interconnection: growing interconnectedness raises the potential for failures to propagate across operational ecosystems.

These call for a shift from a **model-centric** to a **system-centric** governance posture, one that addresses not only outputs but the mechanisms through which decisions are generated, actions executed, dependencies managed, and accountability maintained across the lifecycle.

The Risk Dimensions and the Lifecycle Principle

An **Agent Risk Management (ARM)** framework should build on MRM principles, extending its scope and reach to address the limitations around the four risk dimensions (**Data, Vendor, Architecture, Human Factor**).

The analysis presented in the previous chapter demonstrates that these dimensions remain relevant across all GenAI architectures, while their relative importance evolves as systems progress from stand-alone LLMs to autonomous agents and Multi-Agent Systems. As illustrated in Figure 6, the increasing autonomy and interconnectedness of advanced AI systems progressively transform localized model-level risks into operational and, ultimately, systemic risks. This progression provides

Framework	Scope	Primary anchor
SR 26-2 (Fed, FDIC, OCC; 17 Apr 2026) — <i>supersedes SR 11-7 as the foundational MRM framework</i>	Risk-based, materiality-proportionate MRM for traditional and non-generative models. Expressly excludes generative and agentic AI (fn. 3) , directing institutions to govern such systems through broader risk-management practices.	The MRM baseline must be extended, and the explicit gap must be filled.
SS1/23 (PRA, 2023) updated in 2026	Five principles (model identification & tiering, governance, development/use, independent validation, mitigants); definition reaches complex ML.	Validation, risk-tiering, governance.
EU AI Act (Reg. 2024/1689)	High-risk systems: Art. 9 (continuous lifecycle risk management), Art. 10 (data governance), Art. 13 (transparency), Art. 14 (human oversight), Art. 15 (accuracy, robustness, cybersecurity). Obligations are tiered by risk classification . Foundation models fall under a separate GPAI regime (Art. 51–55).	Data, transparency, human oversight; GPAI → Vendor dimension.
DORA (Reg. 2022/2554)	ICT risk management, ICT third-party risk , operational resilience, incident reporting, resilience testing, oversight of critical providers.	Vendor / dependency risk.

TABLE 3: Overview of Regulatory Frameworks and Their Primary Governance Focus.

the conceptual foundation for ARM: governance mechanisms must evolve in parallel with system complexity, adapting both the nature and intensity of controls to the corresponding risk profile. Consequently, an ARM must be founded on a lifecycle-oriented governance model⁴ in which each risk dimension’s relevance varies across the different stages of system development and operation.

- During design and development, governance primarily addresses data requirements, technology selection, architectural choices, and the definition of operational boundaries.

- During implementation and validation, attention shifts toward execution logic, reasoning quality, tool orchestration, memory mechanisms, and dependency management.
- Once deployed, governance focuses on runtime monitoring, operational resilience, behavioral stability, and the effectiveness of human oversight.
- Periodic reviews then consolidate operational evidence to reassess risk exposure, recalibrate controls, and ensure that governance remains aligned with the system’s evolving characteristics.

⁴“The risk management system shall be understood as a continuous iterative process planned and run throughout the entire lifecycle of a high-risk AI system, requiring regular systematic review and updating.” EU AI Act; Chapter III: High-Risk in AI System; Section 2: Requirements for High-Risk AI Systems; Article 9: Risk Management System.

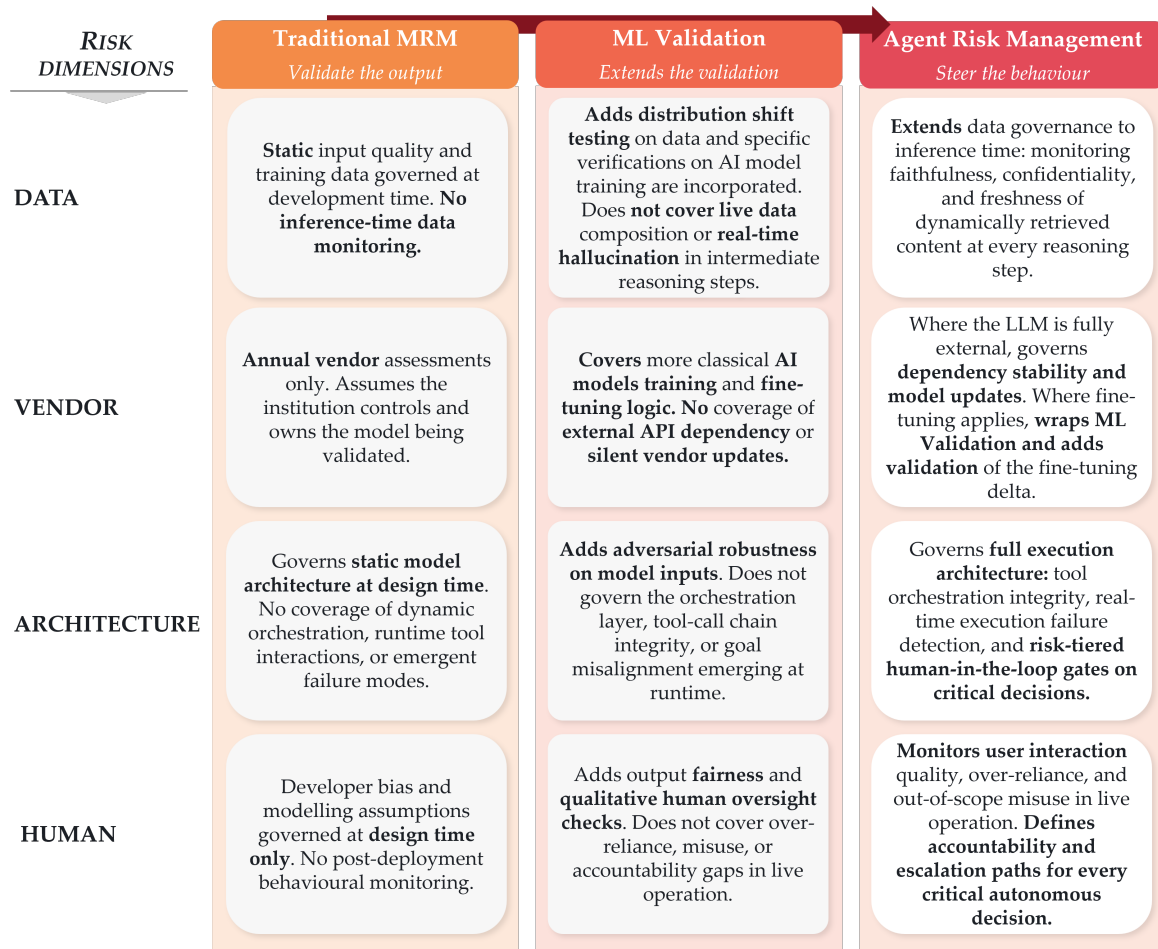


FIGURE 6: The three step evolution of model validation and governance in the AI-era.

This lifecycle perspective represents the principal distinction between ARM and conventional MRM. Whereas traditional model governance primarily evaluates model performance at predefined validation points, ARM recognizes that the risk profile of agentic systems continues to evolve after deployment through interactions with users, external tools, changing environments, and vendor-controlled components. Effective governance therefore requires continuous monitoring and adaptive control mechanisms capable of identifying, assessing, and mitigating risks throughout the entire operational lifecycle.

Accordingly, the objective is straightforward to state and demanding to implement: **higher levels of autonomy must be matched by correspondingly higher levels of transparency, accountability, and control.**

The Pillars of an ARM

ARM operationalises the lifecycle principle of Section “The Risk Dimensions and the Lifecycle Principle” through six complementary components: **an AI inventory, risk-based tiering, incident management and deterministic safeguards, a control toolset, behaviour-level monitoring, and an evaluation-driven lifecycle.** These are not independent controls implemented independently but an integrated framework, aligning governance activity with the evolving risk profile of an autonomous system. The six also divide along an architectural line that matters for deployment. Some must be **embedded in the agent and its runtime**: e.g., logging, tracing, and guardrails should be embedded by design rather than introduced as ex-post controls, and therefore presuppose the control over

the stack argued in Section “*Implementation as a Risk Decision: The Case for a Controllable Stack*”. Others instead are overlaid around the agent, consuming its telemetry to assess and steer behaviour: e.g., risk tiering, Key Risk Indicators, and the evaluation loop all operate on what the embedded layer emits. A functioning ARM requires both: instrumentation without analytical oversight may generate unused data, while oversight without instrumentation may lack an adequate evidentiary basis.

Core Components

The Centralized AI Inventory & Agent Registry (Overlaid)

Most large enterprises lack a comprehensive and consistently maintained inventory of their fragmented AI deployments (across cloud platforms, SaaS tools, and localized developer pilots). However, compliance frameworks like the **EU AI Act** and **ISO/IEC 42001** position inventory tracking and risk classification as absolute foundational requirements for effective governance. To bridge this gap, the **AI Agent Registry** acts as a centralized or federated “central registry” and discovery platform, indexing each agent’s identity, version, and capabilities into standardized formats (e.g., JSON schemas or “agent cards”). It fulfills six core governance functions:

1. **Agent Registration:** ingests metadata payloads (name, version, endpoint URLs, and declared skills) via secure endpoints to standardize deployments;
2. **Discovery and Search:** enables clients, orchestration layers, or other agents to query and locate active agents by capability, tag, or keyword;
3. **Metadata Management:** maintains rich technical logs, including authentication profiles, interaction protocols, data types, and cryptographically verifiable trust credentials;

4. **Uptime Monitoring:** tracks operational uptime via periodic checks to automatically mark offline or unhealthy agents as stale;
5. **Access Control & Role-Based Access Control (RBAC):** enforces granular governance and Access Control Lists (ACLs) to restrict who can register or invoke specific agent capabilities based on client permissions;
6. **Audit Logging & Observability:** feeds metrics, logging, and span-level tracing into central monitoring tools to provide a clear audit trail of every query and API/tool invocation.

Risk-based Tiering (Overlaid)

Agents differ widely in risk exposure; e.g., an informational assistant on public data is a different proposition from an agent with write access to critical systems and regulated data. An ARM must classify agents through a **multi-driver assessment**, capturing dimensions such as the agent’s level of autonomy, the sensitivity of the data it handles, its potential to propagate failures across systems, and its exposure to volatile external dependencies, aggregated into a small number of **governance tiers**. Quantitative scoring has to be supported by expert override and independent challenge by a second line of defence: numbers inform the classification but do not replace judgement. The tier an agent receives then determines its validation, monitoring, testing, and escalation obligations, aligning oversight with actual impact rather than applying it uniformly.

Incident Management & Deterministic Safeguards (Embedded)

For agents empowered with active tool use, informational guardrails are insufficient; the system requires hardcoded operational constraints to manage real-time exceptions and containment:

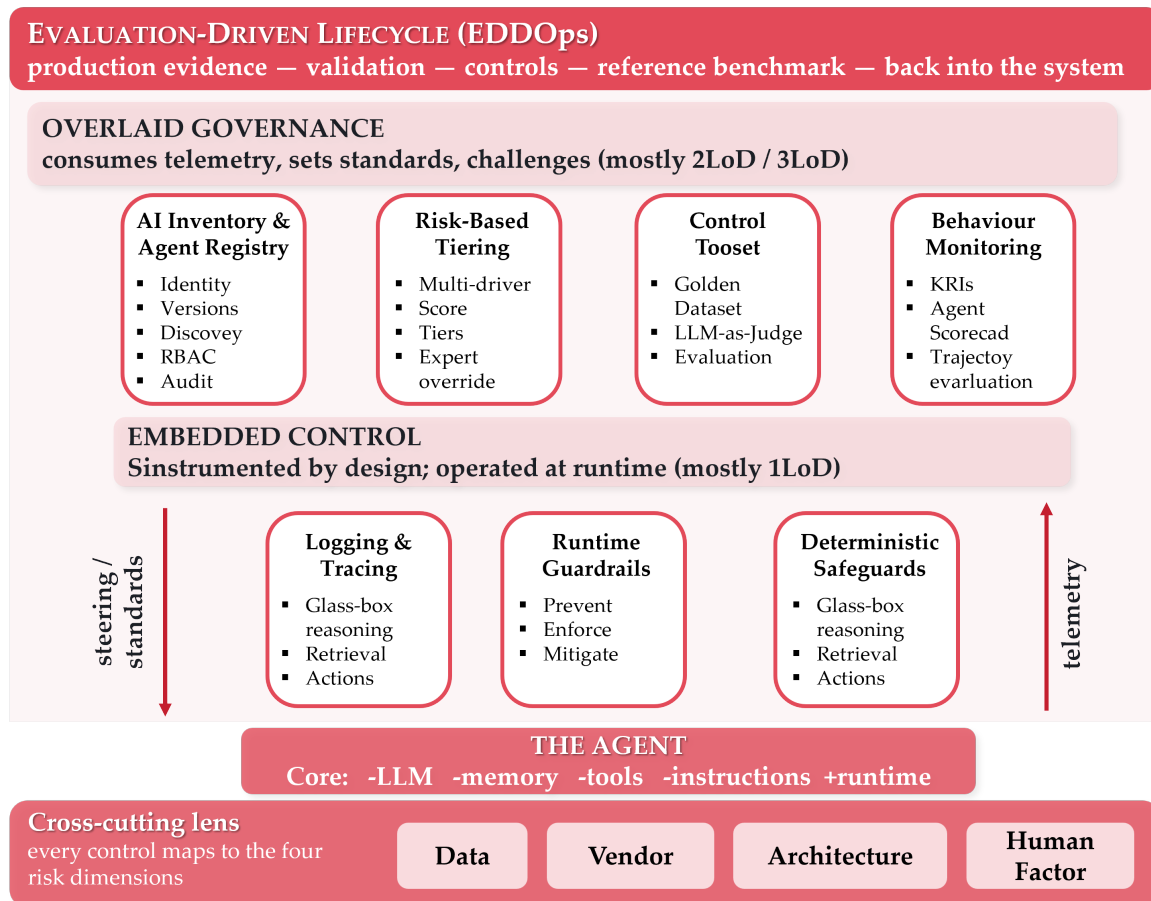


FIGURE 7: EDDOps Governance Framework for Controlled Agentic AI.

- **Granular Access Privileges (Principle of Least Privilege)**: agents must never run with generic corporate system tokens. Each agent must possess a unique, machine-readable cryptographic identity with access permissions restricted to the minimum required tools. Read/write capabilities must be isolated at the infrastructure layer.
- **Automated Kill Switch (Circuit Breaker)**: the runtime environment must be governed by a hardcoded, non-AI programmatic wrapper that continuously evaluates operational telemetry. In the event that an agent enters an infinite planning loop, attempts to invoke an unauthorized API, or breaches quantitative transaction or financial thresholds, this circuit breaker automatically revokes the active session tokens and terminates

execution.

- **State Rollback Mechanisms**: the architecture must enforce event-sourcing or immutable state logging for all tool interactions. If an agent executes a corrupted or unauthorized workflow, operators must have the capability to execute a state restoration to the last verified, structurally sound system state.
- **Agent-Specific Incident Management**: traditional Information Technology Infrastructure Library (ITIL) incident workflows, designed for deterministic software behaviors which may be insufficient to capture stochastic and non-linear failure modes of advanced AI. When a kill switch is triggered or an operational anomaly is flagged, the system must automatically route the event to specialized

triage queues. These queues must preserve an immutable snapshot of the operational state, the complete prompt-trace lineage, and the exact vector index configurations to prevent the recursive propagation of failures across the enterprise ecosystem.

A Control Toolset (Embedded + Overlaid)

Continuous governance requires dedicated infrastructure. ARM combines complementary instruments spanning both layers:

- A **governed reference benchmark** (i.e., a Golden Dataset) capturing not only correct outcomes but the reasoning trajectories experts expect, enabling evaluation of how an agent reaches a conclusion, not only whether it does;
- A **logging and tracing layer** providing end-to-end, glass-box visibility into the agent's reasoning, retrievals, and actions - the evidentiary foundation for explainability, accountability, and audit;
- **Runtime guardrails** spanning preventive controls, operational enforcement, and active mitigation of unsafe behaviour;
- **Automated, reasoning-aware evaluation** that scales assessment of agent behaviour while preserving human accountability within a layered review structure.

Behaviour-level Monitoring (Overlaid)

Because an agent's reliability cannot be inferred from its final output alone, ARM associates each of the four risk dimensions with **Key Risk Indicators** monitoring the conditions that precede failure (e.g., information quality, dependency stability, execution integrity, etc.). Indicators are tailored to the agent's purpose and assessed

at the cadence each requires (e.g., real-time vs weekly), and consolidated into an **Agent Scorecard** that functions as an early-warning governance interface. In parallel, ARM evaluates **trajectories, not just outputs**: it compares observed execution paths against validated operational patterns, assessing the reliability of the reasoning and execution that produced a result, not only the result itself.

An Evaluation-driven Lifecycle

These elements operate as a continuous closed loop. Evidence collected in production (traces, indicator readings, incidents, user feedback) feeds back into validation, controls, and the reference benchmark, so governance evolves alongside the system rather than remaining static after deployment. Governance becomes a continuous capability, not a periodic compliance exercise. Together, these components translate the high-level principles of Section "*The Pillars of an ARM*" into a coherent set of lifecycle controls. Their calibration is inherently risk-based: the specific indicators, thresholds, validation procedures, monitoring frequencies, and supporting technologies are fitted to the institution, the use case, and the agent's assessed risk, and are developed through direct engagement.

Operationalization and Continuous Governance Lifecycle

To transition the ARM framework from conceptual pillars to day-to-day operations, the four core risk dimensions must be mapped directly to clear lines of accountability, explicit audit trails, and formalized enforcement frequencies. Table 4 provides the engineering and risk management teams with a structured checklist of required artifacts to maintain compliance and assure the operational resilience of deployed agentic solutions.

Risk Dimension	Control Objective	Evidence Artefact (Audit Trail)	Accountable Defense Layer	Operational Cadence
Data Risks	Prevent PII leakage in live chats/reports and secure source grounding.	• LLM-as-a-Judge (LLMaaJ) sensitive span flags • Output auditing near-miss trace logs • EU AI Act traceability & audit trail certificates	1LoD: operate data-quality controls; monitor KRIs and triage flagged near-miss traces in real time. 2LoD: set KRI thresholds and grounding standards; independently validate retrieval/generation quality and challenge 1LoD's remediation. 3LoD: assure that the data-control framework is designed and operating effectively, and that the evidence trail is complete.	Real-Time to Daily
Vendor Risks	Control token cost spikes, performance drops, and unannounced model drift.	• Token cost variance data • Production latency logs • Vendor change notifications & drift assessments	1LoD: monitor cost/latency/availability KRIs and raise drift alerts. 2LoD: set drift-tolerance and cost thresholds; independently re-assess model behaviour on drift signals and approve/hold vendor updates. 3LoD: assure that the vendor-oversight and change-control process exists and was followed.	Monthly
Architecture Risks	Suppress prompt injections, execution loop issues, and failed tool invocations.	• Automated guardrail exception logs • Tool call success metrics • Systemic propagation owner-assignment logs	1LoD: operate guardrails and kill-switch; monitor execution KRIs and remediate borderline traces. 2LoD: set guardrail and escalation standards; independently challenge whether the control framework adequately covers injection, looping, and cascading failure. 3LoD: assure the guardrail and owner-assignment processes are designed and operating effectively.	Real-Time to Daily
Human Risks	Prevent automation bias, out-of-scope misuse, and HITL gate bypasses.	• HITL gate verification stamps • Escalation playbook invocation logs • Unowned incident counters & misuse scores	1LoD: execute HITL decision gates and stopper protocols; own escalation and remediation. 2LoD: set oversight standards and KRI thresholds; independently monitor gate adherence and challenge accountability gaps. 3LoD: assure that human-oversight controls are defined, owned, and functioning as intended.	Weekly/ Monthly

TABLE 4: The Operational Governance Matrix - Consistent with the three-lines-of-defence model: the first line owns and operates controls, the second sets standards and independently challenges, and the third provides independent assurance over both.

Conclusions

Agentic AI marks a structural change in the way artificial intelligence enters financial institutions. The shift is no longer limited to systems that generate content or support human analysis: agents perceive context, reason over possible courses of action, use tools, and act within operational environments. As a consequence, the risk profile also changes. Failures are no longer confined to inaccurate outputs; they may propagate through processes, systems, third-party dependencies, customer interactions, and regulatory obligations.

Autonomy cannot be governed as an extension of traditional model risk alone. Classical Model Risk Management remains a necessary foundation, but it was designed around relatively stable models. Agentic systems require a broader governance layer capable of observing and controlling behaviour across the full execution trajectory: how information is retrieved, how reasoning and planning are performed, how tools are invoked, how actions are executed, and how human accountability is preserved. This is precisely the gap that the current supervisory landscape makes visible.

For this reason, **implementation choices** become **governance choices**. Institutions that build on opaque, externally managed models may obtain rapid access to capability, but they also accept limitations in reproducibility, auditability, continuity, data control, and effective challenge. Conversely, a controllable stack, based where appropriate on open-weight models, institution-owned deployment, curated domain-specific data, fine-tuning, version control, and validated tool integration, provides the technical foundation for governable autonomy. The central consideration is therefore not whether the most powerful frontier model can be accessed, but whether the selected architecture is sufficiently accurate, secure, reproducible, explainable, and controllable for the specific use case and risk tier.

These findings have several implications for financial institutions. Agentic AI should not be adopted first and governed later. Before scaling agents into material processes, institutions should identify the agentic use cases already emerging across the organization, classify them by autonomy and impact, decide where a fully-controllable stack is required, define minimum standards for data, vendor, architecture, and human oversight, and establish an ARM framework able to monitor behavior continuously after deployment. Institutions that establish governance capabilities alongside the deployment of agentic AI will be better positioned to scale these technologies while maintaining operational resilience, regulatory compliance, and stakeholder trust.



References

- [1] **Bandi, A. et al.** *The Rise of Agentic AI: A Review of Definitions, Frameworks, Architectures, Applications, Evaluation Metrics and Challenges* September 2025.
- [2] **Board of Governors of the Federal Reserve System, FDIC and OCC.** *SR 26-2: Supervisory Guidance on Model Risk Management.* 17 April 2026.
- [3] **Esposito, D. and Carminati, M.** *Machine Learning Framework.* Just in Time, October 2022.
- [4] **Esposito, D., Musto, M. and Travenzoli, T.** *Model Risk Management of GenAI Workflows.* Just in Time, February 2025.
- [5] **Esposito, D. and Travenzoli, T.** *GenAI Model Risk Management and Governance in Financial Services – From Principles to Practice.* Just in Time, February 2026.
- [6] **European Parliament and the Council.** *Regulation (EU) 2024/1689, Artificial Intelligence Act.* Official Journal of the European Union, 2024.
- [7] **European Parliament and the Council.** *Regulation (EU) 2022/2554 on digital operational resilience for the financial sector (DORA).* Official Journal of the European Union, December 2022.
- [8] **Financial Stability Board.** *Sound Practices for Responsible Adoption of Artificial Intelligence (AI), Consultation Report.* 10 June 2026.
- [9] **Martucci, S. and Ranieri, M.** *AI Risk Management Frameworks.* Research Paper Series N.81, November 2025.
- [10] **National Institute of Standards and Technology.** *Artificial Intelligence Risk Management Framework (AI RMF 1.0).* January 2023.
- [11] **OECD.** *AI Principles* May 2019.
- [12] **Patil, S. G. et al.** *The Berkeley Function Calling Leaderboard (BFCL): From Tool Use to Agentic Evaluation of Large Language Models.* ICML, 2025.
- [13] **Prudential Regulation Authority.** *SS1/23 – Model risk management principles for banks.* Supervisory Statement 1/23, effective from 23 April 2026.
- [14] **Ranieri, M.** *AI Agents and Risk Management.* Just in Time, April 2026.
- [15] **Russell, S. and Norvig, P.** *Artificial Intelligence: A Modern Approach* (4th ed.). Pearson, 2021. For a recent reframing see Bandi A. et al., *The Rise of Agentic AI: A Review of Definitions, Frameworks, Architectures, Applications, Evaluation Metrics, and Challenges*, Future Internet, Vol. 17, No. 9, 2025.
- [16] **Yao, S. and et al.** *ReAct: Synergizing Reasoning and Acting in Language Models.* Princeton University & Google Research. International Conference on Learning Representations (ICLR), 2023.

Sitography

[17] **Artificial Analysis.** *Website.*

[18] **Artificial Analysis.** *Website.*

[19] **Epoch AI.** *Website.*



Innovation

**Classifying Financial
Transaction for
Invoice Matching**

About the Authors

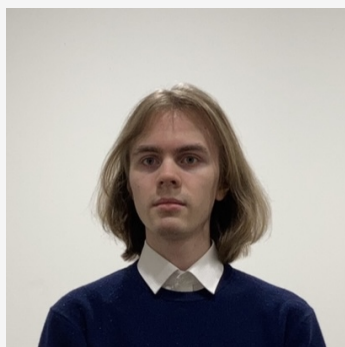


Massimiliano Zanoni:

Senior Manager

With over 20 years experience in the management of complex projects in credit and financial risk he has worked within leading national and international financial services, maturing a combination of quantitative analysis and organizational skills. His consulting experience ranges from model development to the implementation of management reporting system; from to compliance and validation of IRB credit risk framework to setup of ALM systems. He has worked for Veneto Banca as the Head of Methodology and Risk Management Reporting. In the same period, he was also the delegate in the Risk committees of foreign banks and the Risk Management delegate in the Board of Banca Apulia where he directly oversaw the transition process to the Solvency II framework of the participated insurance company. In iason, he has been involved in the development of credit risk solutions and the implementation of a Dynamic Balance Sheet framework.





Evgenii Veksin:

Analyst

He holds bachelor's degree in economics at Higher School of Economics (Moscow). Secondly, he holds a master's degree in Management Engineering at Politecnico di Milano. In Iason, he was involved on the Credit Risk topics: Stress Test project, internal GRisk models.



This work addresses the automated classification of bank transactions to facilitate the reconciliation of invoices, a critical and often tedious task in the financial control process. The study evaluates two primary methodologies: a Machine Learning approach known as K-Nearest Neighbors (k-NN), based on the Levenshtein distance, and a modern Large Language Model (LLM) approach based on the BERT architecture. To manage computational efficiency and scalability of the Machine Learning approach, the work explores Hierarchical Agglomerative Clustering (HAC) and OPTICS, which successfully reduced the training dataset size by 5 to 6 times while maintaining an accuracy level of approximately 99%. The LLM component utilizes a hybrid framework combining BERT embeddings with Bag-of-Words (BoW) representations to effectively handle low-data environments and the specific nuances of financial jargon and abbreviations. Experimental results for both methodologies demonstrate accuracy exceeding 99%, rivaling traditional manual rule-based approaches. Furthermore, the study introduces confidence metrics, derived from average neighbor distances in the k-NN model or probability distributions in the LLM approach, to selectively spot "doubtful" cases for manual review. In conclusion, the integration of these AI and ML techniques provides a precise, consistent, and scalable solution, allowing accountants to focus on higher-value analytical activities.

THE process of reconciling invoices with corresponding bank transactions serves as a fundamental control mechanism to ensure consistency between a company's internal financial records and the official statements issued by its banking institutions. This reconciliation is essential to guarantee that accounts payable and accounts receivable accurately represent the underlying cash movements recorded by the bank. This can be challenging and time-consuming and has introduced areas for automated support across invoice processing tasks and research provides evidence that the use of hierarchical information can highly improve classification precision [9].

The effectiveness of this procedure depends critically on the correct classification of bank transactions, as misclassifications can compromise the automation of the matching process and lead to discrepancies, manual interventions, and potential accounting errors. More specifically, the matching activity involves the systematic analysis of a large amount of structured and unstructured data, in particular:

- **Invoices recorded in the company's accounting system**, representing the

formal documentation of outstanding obligations to suppliers (accounts payable) and amounts due from customers (accounts receivable). These invoices reflect the organization's official financial commitments and claims⁵.

- **Bank statement lines**, i.e. the official bank transaction descriptions recorded by the bank, which can be collected from an external servicer, upon authorization.

The precise synchronization between expected and actual payment datasets facilitates the automated identification of financial correspondences. Within this framework, the granularity and accuracy of bank transaction classification emerge as critical prerequisites for improving the reliability, efficiency, and scalability of automated invoice matching.

High-performance classification tools are essential for isolating invoice-related entries from extraneous operational "noise"—such as commissions, taxes, or interest charges. By effectively filtering these datasets, such tools significantly constrain the search space for potential matches, thereby stream-

⁵In the Italian context, where invoices are digitally recorded by the Internal Revenue Service, these can be downloaded from "Cassetto Fiscale", after the firm has authorized the servicer.

lining the reconciliation process and reducing the risk of false-positive correspondences. With increasing data volumes and the demand for rapid and accurate decision-making, the integration of AI and ML has become pivotal. [2] Specifically, the application of machine learning has the potential to provide novel insights into empirical data and to improve predictive performance. [7] This study focuses on the use of AI in the classification of bank transactions. Through the application of Machine Learning (ML) methods and Large Language Model (LLM) techniques, it explores the benefits and challenges of adopting these technologies, focusing on their potential to enhance efficiency and accuracy within management accounting as in [2], and shows that AI-based systems supplement human skills, allowing accountants to focus on higher-value jobs as in [13]. The study is organized as follows:

- **Chapter 1** examines a K-nearest neighbors (KNN) model based on the Levenshtein distance, to extract meaningful information from transaction descriptions and classify the transactions under analysis. The chapter further evaluates two strategies for reducing and optimizing the training set, by statistically aggregating similar descriptions into homogeneous clusters from which representative points are selected to construct a reduced training set. The performance of the KNN classifier is then assessed across multiple reduced training configurations, to determine the extent to which the process can serve as a viable approach when the size of the transactions set grows significantly.
- **Chapter 2** extends the analysis by introducing modern LLM-based techniques, specifically those derived from the BERT architecture, to enhance traditional transaction classification methods. In contrast to classical models relying on surface-level text matching, advanced natural lan-

guage processing (NLP) algorithms can capture the contextual meaning of textual components, enabling more nuanced interpretations. This capability is particularly relevant for financial transactions, where the sequence of keywords may play a decisive role in determining the correct classification.

Dataset and Rule-based Classification

The dataset includes more than 800 bank transactions relative to one quarter activity from the bank accounts of the selected firm. Such a dataset clearly is not suitable for big data analysis, but it may be adequate to train a ML model in order to recognize existing text structures and possibly interpret new descriptions.

The description of account transactions is clearly bank specific and similar transactions may present different descriptions, i.e. the text may or may-not include the counterpart name, the invoice number, the reference date or other technical codes. In addition, the structure of descriptions is highly dependent on the category of the transaction. Descriptions of technical transactions, like commissions or interests, are typically rather homogeneous in comparison to business transactions, for example the following descriptions represent the same type of business transaction:

- BEU INTERN BANK Bonifico da Voi disposto a favore di;
- BON.UE CAN.TELEM. Bonifico da Voi disposto a favore di.

At first, transactions were classified through manually defined rules, based on defined keywords: each time a set of keywords was found in the description field, the transaction was classified accordingly. The following Table 5 lists a sample of keywords and their relative rule-based classification.

This rule-based approach is simple to understand and works reasonably well, but it

Keywords in Description	Transaction Class	Transaction Type
"ACCR.IMP."	Pending Invoice	Total
"ANT FATT O DOC ITA"	Pending Invoice	Pending
"ACCR.BEU", "BONIFICO A VOSTRO FAVORE", "BONIFICO A VS.FAV."	Transfer	Credit
"BEU", "BON.UE", "BONIF."	Transfer	Charge
"PAGAMENTO ADUE"	Transfer	Payment Order
"DISP.BEU STIP."	Transfer	Salary
"STORNO"	Transfer	Reversal Movement
.....		

TABLE 5: *Keyword Based Association.*

lacks flexibility and requires a considerable effort. Each new bank is associated with a specific set of manually defined keywords and rules whose assignment is considered as the "correct" reference. Creating and maintaining these training sets involves a significant manual effort, tailored to each new bank, and requires updates whenever a bank changes the format or style of its transaction descriptions. To automatize the identification of keywords and the classification of transactions, a Machine Learning approach, based on K-Nearest Neighbors and Levenshtein distance, was tested with surprising results. The following paragraph describes the methodology adopted.

The K-Nearest Neighbors Approach

This chapter introduces the k-Nearest Neighbors algorithm, a fundamental technique in machine learning that is widely applied to classification problems, including those involving textual data. The central idea behind k-NN is straightforward: when a new input sample is presented, the algorithm identifies the k training instances that are most similar to it, according to a chosen distance metric. The class of the new sample is then inferred from the majority class among these nearest neighbors.

In text classification tasks, the notion of "distance" is typically interpreted as a measure of linguistic similarity. One common approach is to use the **Levenshtein distance**, which calculates the minimum number of character-level edits or insertions, deletions or substitutions, required to transform one string into another. This metric effectively captures how similar two short text strings are, making it suitable for tasks such as matching textual labels or categorizing brief messages.

Formally, k -NN is a **non-parametric, instance-based supervised learning algorithm**. Unlike parametric models, it does not build an explicit predictive model during training. Instead, it simply stores the training data and defers all computations to the prediction phase. When classifying a new instance, the algorithm relies directly on the stored data to determine the appropriate class label.

Due to its simplicity, transparency, and ease of implementation, k-NN is frequently used as a **baseline method** in natural language processing applications, particularly those involving the classification or clustering of short text segments.

The k-Nearest Neighbors algorithm works by finding the closest number (k) of training samples to the new sentence which needs to be classified, in terms of the defined distance. As any supervised it can be split into

two phases:

- **Training Phase:** in contrast to model based algorithms, k-NN does not involve parameter estimation nor gradient-based optimization. The training phase consists solely of storing the entire dataset of labeled transaction descriptions. Each description retains its assigned class and serves as a reference point for subsequent comparisons.
- **Prediction Phase:** when a new, unclassified transaction description is provided, the algorithm computes its distance to every stored training instance. It then identifies the k closest descriptions, those exhibiting the greatest textual similarity according to the selected metric. The class most frequently represented among these k neighbors is subsequently assigned to the new description. This majority-voting mechanism enables k-NN to leverage local patterns in the data, making it particularly effective when similar descriptions consistently correspond to the same transaction category.

By relying directly on the structure of the training samples and their linguistic proximity, the k-NN algorithm provides an interpretable and flexible baseline method for transaction classification. Its ability to generalize from local patterns makes it especially well-suited for environments where descriptions are heterogeneous yet recurrent, such as bank statements.

Through these two phases, the k-NN algorithm provides a straightforward yet powerful framework for transactions classification, laying the groundwork for more advanced techniques explored in the following sections. This chapter is organized as follows:

A. Introducing the Levenshtein Distance with k-NN;

B. Optimizing the Training Dataset;

C. Presentation of Results;

D. Conclusions.

Introducing the Levenshtein Distance with k-NN

The key to applying k-NN to text classification, is to define a robust measure of the difference between two sentences. In this work the Levenshtein distance was adopted for this purpose. It provides the difference between two sentences in terms of the number of single-character edits (insertions, deletions, or substitutions) required to change one string into the other. When using this methodology, a strong assumption is made, i.e. descriptions that are close are also semantically similar, hence they can be classified with the same tag.

A refinement of the standard Levenshtein distance is its normalized form, calculated as the standard distance (number of edits) divided by the length of the longer string. By definition, this measure ranges between 0 and 1, i.e: it is 0 if the two strings are identical and 1 if every character has to be changed to obtain the test description. The classification process is described by the following steps:

- 1 Select the training Dataset, i.e. the list of transaction description classified with a deterministic rule-based approach (example in Table 5).
- 2 Chose the test sentence, i.e. a new transaction description to be classified.
- 3 Calculate the distances between the new description and every single sentence in the training dataset. One of the following different cases may occur:
 - a) The two sentences belong to the same category, "accr.imp. non ant. cod. disposiz: saldo fattura

n. 11". " accr.imp. non ant. cod. disposiz: saldo fattura n. 22". Levenshtein dist. = 2; Normalized dist. = 0.037.

- b) The two sentences partially overlap but belong to different categories, "accr.imp. non ant. cod. disposiz: saldo fattura n. 11", " accr. beu cod. disp.: cash bon.eu" the Levenshtein dist. = 31 , Normalized dist. = 0.58.
- c) The two sentences have no similarity thus belong to different categories, "accr.imp. non ant. cod. disposiz: saldo fattura n. 11", "pagamento rata mutuo/prestito pagamento rata n. 9 del finanzia") the Levenshtein dist. = 51 , Normalized dist. = 0.83.

- 4 Find the closest neighbors, i.e. the closest k sentences in the training set, those having the lowest Levenshtein distance to the test sentence.
- 5 Assign the class; the most frequent tag found among the k-nearest neighbors is selected as the correct label to be assigned to the test sentence. In the case of K=3, the tag selected is that of the majority among neighbors.

Advantages and disadvantages of the k-NN approach are summarized in Table 7.

The first two disadvantages are successfully addressed in this work, as described in the following paragraphs. The third disadvantage is more relevant in the context considered, however, it should be efficiently addressed with the the LLM approach analyzed in the next chapter.

The training dataset may include transactions with rare classifications or even occurring only once (e.g , "bonifico storno", see Table 8). Clearly, such cases cannot be classified correctly with the current approach since there are not enough neighbors with the same category to be selected and determine the correct classification. Therefore all transactions with classifications occurring

only once or twice in the dataset will be excluded from the test, but still considered in the training set.

The Levenshtein distance of descriptions close to the transaction in analysis is improved by eliminating from the text irrelevant components, which generate noise, as described by the following procedure:

- Convert every text to lower case, as the case type does not change the meaning of keywords.
- Delete large meaningless sequences, not useful for classification purposes but adding noise to the analysis, such as strings containing at least 6 digits. Example is presented in line (1) of Table 6.
- Truncate descriptions, usually key information is located in the first half of the text, whereas the second half includes transaction specific data, such as counterparty name, invoice number, etc. not relevant for classification. This approach reduces the noise but is mainly valid for long description of the type presented in the line (2) of table 6.

The three lines in Table 6 provide an example of the effect of applying the previous rules.

Procedure Description

To maximise the size of the training set used in this exercise, a leave-one-out validation approach was adopted. Under this procedure, each transaction is tested individually using all remaining transactions as the training set. Concretely, for each of the 836 original descriptions, one description is removed and treated as the test instance. The Levenshtein distance is then computed between this test description and the remaining 835 transactions to identify the three closest matches. The test transaction is subsequently assigned to the most frequently occurring class among these three (i.e., the class appearing in at least two out of the three cases).

Before	After
(1) bon.ue can.telem. 0124092836794514 w0242102615507532400000272	bon.ue can.telem.
(2) bonifico a vostro favore bonifico sepa da banca popolare di puglia e Basilicata società cooperativa per azioni per a saldo vostra fattura 144 20/12/202 trn	bonifico a vostro favore bonifico sepa da banca popolare di puglia e Basilica
(3) ACCR. BEU COD. DISP.: 0124092731001227 CASH BON.EUR.UNICO VS.FAV. CON CONTAB. NOTPROVIDED 78 20240730	accr. beu cod. disp.: cash bon.eur.un

TABLE 6: Examples of Description Cleaning.

Advantages	Disadvantages
1. <u>Minimal training</u> : the "training" simply consists in storing the data, making it very fast to "train" the model. The computational work is all done at prediction time (i.e. measuring distances).	1. <u>Computational Cost</u> : calculating the Levenshtein distance between a new sentence and every single sentence in a large dataset is computationally burdening. This is a major drawback for real-time applications with a high volume of data.
2. <u>Simple and Intuitive</u> : the approach represents a natural way to think about similarity, the more alike two strings are, the closer their distance.	2. <u>Scalability Issues</u> : as the training dataset grows, the prediction time increases linearly, making it a poor choice for large-scale production systems.
3. <u>Good for small datasets</u> : this method can perform well with small datasets where the relationships between sentences are not too complex.	3. <u>Overemphasis on Character Order</u> : Sentences that are semantically very similar might have a high Levenshtein distance due to different word choices or sentence structures. For example, "How do I turn on the light?" and "Can you switch the light on?" have a high Levenshtein distance but a similar meaning.

TABLE 7: k-NN Advantages and Disadvantages.

If all three neighbours belong to different classes, the classification is determined by the single transaction with the smallest distance to the test description. As noted previously, transactions associated with labels that appear only once or twice cannot be included in the test set, since they cannot be meaningfully evaluated within the leave-one-out procedure. In this dataset, only two such transactions exist (see Table 8),

each appearing only once. Consequently, 834 transactions remain eligible for testing under the leave-one-out framework. Comparing each test description against all others becomes computationally burdensome in large datasets, and the efficiency of the process decreases as the training set grows. Given that many descriptions are highly similar, computing the distance to each individual instance is suboptimal. A potential

Full Description	Truncated Description	Classification
"PAGAM. TRAMITE POS COOP.RADIO TAXI PADOVA VIA 13/05-14:29 - Carta n.5259XXXX XXXX XX07ABI: 09517 COD.3058793/000165"	"pagam. tramite pos coop.radio taxi padova via13/0"	None
"STORNO BEU STORNO DEL BONIFICO RIF.71 – 0125011037593015 CONTENUTO NEL FLUSSO RIF.DICEMBRE 2024"	"storno beu storno delbonifico rif.71 -"	Bonifico storno

TABLE 8: Rarely Occuring Transactions.

improvement would involve reducing the effective training set by clustering similar descriptions and subsequently computing distances only against representative elements of each cluster rather than against all individual transactions. This shifts the challenge from calculating a full pairwise distance matrix to developing a robust and statistically sound method for aggregating similar transactions into coherent groups, ideally containing transactions belonging to the same classification. The following paragraphs describe and compare two different grouping methodologies.

Optimizing the Training Dataset

Many transaction descriptions differ only for minor details which, for the purpose of classification, can be neglected. This means that transactions could be grouped into homogeneous bundles, to reduce computation effort and time, maintaining the information necessary to correctly classify them. The structure of the dataset is more effectively visualized using the Multidimensional Scaling (MDS) algorithm, which provides a geometric representation of the data by taking as input the matrix of pairwise Levenshtein distances and projecting the observations into an abstract Cartesian space (see Figure 8). The objective of MDS is to preserve the original similarities between data points: transactions that are highly similar in their textual description (i.e., with

small Levenshtein distances) are mapped to nearby locations, whereas dissimilar transactions are placed farther apart. Through this transformation, recurring structures in data, such as clusters, gradients, and anomalies, become visually apparent, even though they may be difficult to detect when examining the raw textual data directly. This representation allows:

- To highlight the relative ‘density’ of groups, showing which classes of transactions are numerous;
- To mark isolated (i.e. exceptional) transactions which may signal anomalies or simply outliers.

To create this homogeneous samples, two different clustering approaches are tested, both based on the Levenshtein Distance Matrix, listed together with the sampling procedure:

- a. Hierarchical Agglomerative;
- b. OPTICS clustering;
- c. Sampling Procedure.

a. Hierarchical Agglomerative Clustering
Hierarchical Agglomerative Clustering (HAC) is a bottom-up clustering methodology that allows to construct a tree-like hierarchical structure by iteratively merging data points or clusters based on their

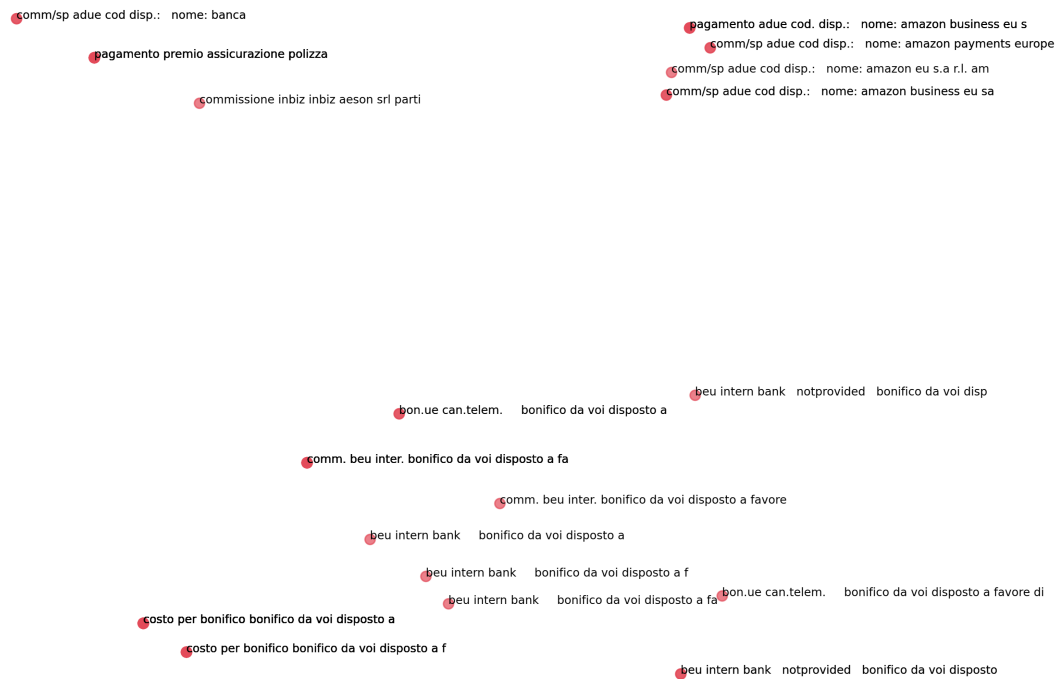


FIGURE 8: MDS Visualization of a Portion of the Transactions Dataset.

pairwise similarity. The process begins by treating each observation as an individual cluster. At each successive step, the two clusters exhibiting the smallest dissimilarity, typically defined by a chosen distance metric, in this study the Levenshtein distance, are merged to form a new, larger cluster. This iterative aggregation continues until all observations have been combined into a single cluster, thereby producing a complete hierarchical representation of the dataset. This approach is particularly effective for text-based similarity analyses because it makes the underlying structure of the data explicit. HAC captures similarity relationships at multiple levels of granularity: initially, highly similar descriptions are grouped into small, tight clusters; subsequently, these clusters are merged into progressively broader groupings. The resulting hierarchy is visualized through a **dendrogram** (see Figure 9), which provides an interpretable graphical summary of the clustering process, with the horizontal axis representing the distance at which clusters are merged, while individual descriptions are arranged along the vertical axis. Branching points indicate cluster separations, and

the height of each branch reflects the dissimilarity between the merged clusters. A key advantage of HAC is that it allows the analyst to determine the desired cluster granularity. Selecting a particular height at which to “cut” the dendrogram, is equivalent to setting a distance threshold, hence the clustering process can be stopped at the desired level. A lower threshold yields a larger number of small, highly homogeneous clusters, whereas a higher threshold produces fewer, broader clusters. Selecting the optimal threshold is inherently subjective: the literature recognizes that no universal rule exists for determining an optimal cut-off, and that the choice depends heavily on the domain, the distance metric, and the overall analytical objective (Sebastiani P. and Perls, T., 2016 [12]).

In this study, the threshold was selected to create the smallest number of clusters that still contained transactions with coherent and meaningful classifications. The thresholds chosen were 4 for the raw Levenshtein distance and 10% (i.e., 0.1) for the normalized version; transactions with a distance below these respective thresholds were grouped together, reflecting the

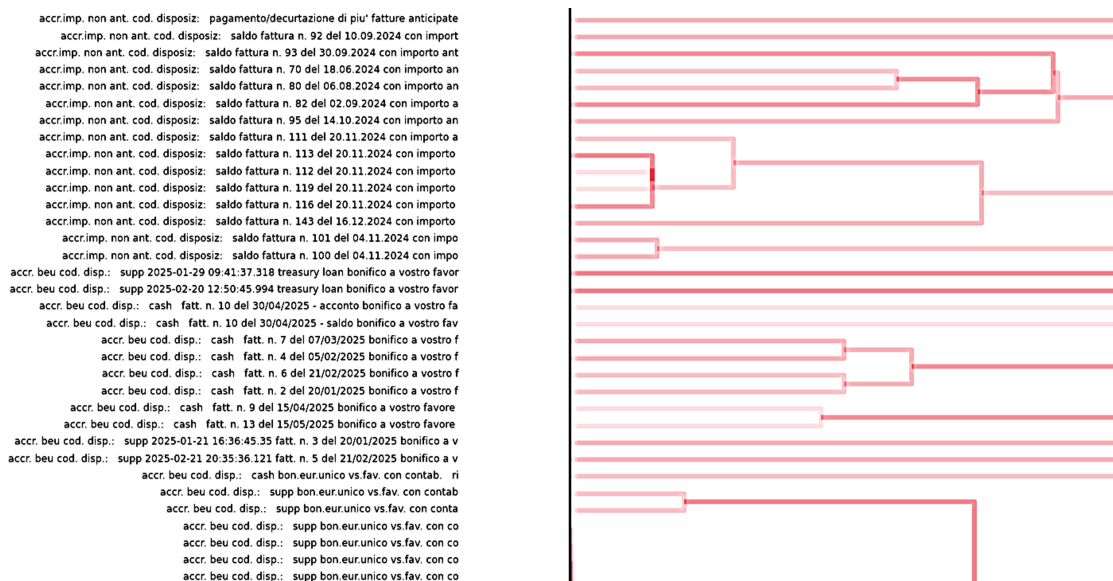


FIGURE 9: Dendrogram Visualization of a Portion of the Transactions Dataset.

assumption that such descriptions are sufficiently similar to be considered part of the same cluster.

Overall, HAC provides a flexible and interpretable framework to examine structural patterns in text-based datasets, enabling both fine-grained and high-level insights into similarity relationships among transaction descriptions.

b. OPTICS Clustering

The second grouping approach considered in this study, is the OPTICS algorithm (Ordering Points To Identify the Clustering Structure). This is a primary density-based methodology for database mining, which groups the points in "regions" with high density and defines isolated points as "noise". It does not produce a clustering of a data set explicitly; but instead it creates an augmented ordering of the database representing its density-based clustering structure. [3] It was specifically designed to identify clusters in datasets characterized by heterogeneous density distributions. This capability makes it particularly well suited to the present application, where groups of string descriptions may vary substantially in length and complexity. Unlike algorithms that assume a uniform density across clusters, OPTICS constructs a plot

that captures the intrinsic clustering structure at multiple density levels. As a result, it can effectively detect clusters composed of either short or long strings, accommodating the natural variability in string-based distance measures.

A notable advantage of OPTICS is its parsimonious parameterization. The primary parameter to be specified is the minimum number of points required to form a cluster, which controls the granularity and stability of the resulting structure. Conceptually, this parameter plays a role analogous to the distance threshold in hierarchical clustering, as both influence the level at which clusters emerge. In the context of this exercise, setting this parameter is straightforward, it is defined to match the value of K , (i.e. 3) to ensure that clusters consist of at least three related descriptions while maintaining sensitivity to meaningful local density variations.

c. Sampling Procedure

The distribution of data points across groups is markedly uneven. A substantial proportion of transactions is concentrated within a limited number of large groups. Several of these groups contain more than one hundred transactions and are characterized by highly recurrent and standardized

descriptions — such as “*commissions*” or “*fiscal transfers*” — which appear frequently throughout the dataset. Beyond these dominant groups, the dataset comprises numerous smaller clusters of varying sizes. Finally, a residual set of points remains that cannot be meaningfully assigned to any cluster. These observations are treated as *outliers* due to their idiosyncratic characteristics and low frequency.

Once transactions have been clustered based on their pairwise distances, three points per group are drawn from each group to construct the reduced training set, an arbitrary choice aligned with the parameter k . To avoid introducing additional noise into the subsequent classification process, it is essential that the selected points faithfully represent the predominant transactional behaviour within each group. For this reason, the sampled points are drawn exclusively from the *most frequent* transaction class within each cluster.

The **treatment of outliers** within this sampling procedure must be handled with particular care. While selecting representatives from groups of normal size is straightforward, the handling of points that cannot be aggregated into meaningful clusters differs substantially between clustering methods:

- With hierarchical clustering outliers tend to be aggregated into very small, isolated clusters, ultimately yielding “micro-clusters” of one or very few points. This occurs because the chosen distance threshold prevents them from merging with any larger group.
- On the other hand, OPTICS directly classifies such observations as *noise*, since these points fail to form a valid cluster with the required minimum size of three elements. In this case, therefore, outliers appear explicitly as points that do not satisfy the density conditions necessary for cluster membership.

The question of whether to include these

outliers in the representative training sample is non-trivial. On the one hand, outliers generally correspond to rare transaction types, meaning that their inclusion may enhance the model’s ability to classify infrequent but potentially important cases. On the other hand, their heterogeneity and scarcity may introduce noise or reduce the quality of the training set. To assess this trade-off empirically, two alternative training samples were constructed and evaluated: one including the outliers identified by the clustering process, and one without them. This dual approach enables a clearer understanding of the impact that rare and irregular transactions have on the overall model performance. The results of the test are presented in the next section.

Test Results and Conclusions

This paragraph presents the results of the exercise highlighting the accuracy of the different sampling approaches, compared to the initial rule-based classification, considered as the correct classification.

The three methods are summarized in the following paragraphs:

- A. The **direct K-NN algorithm**, the full dataset is used as the reference set, i.e. each transaction is compared with all remaining points;
- B. The **Hierarchical algorithm**, transactions are sampled, each test point is compared only with three points from each sample;
- C. The **OPTICS algorithm**, transactions are sampled differently but points are selected as in option B.

Within each sample approach classification is based on the k-NN algorithm using the Levenshtein distance, but the leave-one-out is applied only to option A as, in the others, grouping and the relative sampling is performed once before testing. In case B and C, when the point to be tested belongs to those selected from the clusters, i.e. it

	Simple k-NN	Hierarchical Agglomerative	OPTICS
	Leave-one-out	Clustering	Clustering
Training set	834(833)	189(188)	177(176)
Outliers #	-	69	28
		Performance with outliers	
Mistakes	6/834	5/834	8/834
Accuracy	99.3%	99.5%	99%
Errors below threshold	-	0/5	2/8
		Performance without outliers	
Training set	Outliers not defined for this method	120(119)	149(148)
Mistakes		9/834	6/834
Accuracy		99%	99.3%
Errors below threshold		0/9	2/8

TABLE 9: Summary of Test Results.

is included in the reduced training set, it will be excluded from the training set for this particular case. This process can be easily performed with k-NN, where no pre-determined model is defined.

As summarized in the Table 9, with Hierarchical Clustering the size of the final sample reduces to 189 from the original 834 (a 77.4% reduction), while with OPTICS it reduces to 177 including outliers (a 78.7% reduction).

If outliers are excluded instead, the sample reduces to 120 points with Hierarchical Clustering (a 85.5% reduction) and to 149 with OPTICS (a 82.1% reduction) but, in both approaches, the samples size also depends on the choice of parameters. These four samples, with and without outliers, are used as training sets for the classification exercise with the usual k-NN methodology. As shown in Table 9, the performance are practically the same for both methods with or without outliers, and mirror the results

obtained from the Simple k-NN exercise. This proves, beyond any doubt, that the grouping procedure maintains the same good classification performances while significantly reducing the size of the training set.

A detailed analysis of errors can provide information on the boundary between "confident" and "doubtful" classifications and define a rule-of-thumb to activate a direct manual control of doubtful classifications.

Classification Errors

Looking at the classification errors in tables 10 and 11, for each of them we can notice that the average distance of the 3 nearest points is quite higher than for descriptions correctly classified.

Correct classifications present an average distance to their 3 nearest neighbors of about 4.5% for OPTICS, which decreases to about 2.5% for the simple k-NN, while, in case of errors this average increases sig-

Test ID	Test Description	Real Label	Assigned Label	Top 3 ID	Average Distance
84	bonifico a vs.fav. saldo ft 7 del 04.04.25 cro: bonifico n. disposto	transfer, credit	transfer, charge	[2, 89, 204]	61.2%
202	liq. inter. ant.f competen	interests/charges	commissions	[603, 12, 70]	46.4%
472	comm/sp adue cod. nome:banca	commissions	transfer, payment order	[727, 471, 598]	25.7%
601	comm/sp adue cod. nome:banca	commissions	transfer, payment order	[727, 471, 598]	25.7%
603	liq. inter. ant.f competen	interests/charges	commissions	[202, 12, 70]	46.4%
704	bonifico a vostro favore bonifico sepa da banca popolare di puglia e basilica t	transfer, credit	commissions	[719, 23, 37]	52.9%
706	mav int.	external services	tax	[107, 105, 55]	55.6%
719	bonifico a vostro favore bonifico sepa da banca pop. alto adige per	transfer, credit	commissions	[704, 23, 37]	51.7%

TABLE 10: OPTICS - Error list with outliers.

Test ID	Test Description	Real Label	Assigned Label	Top 3 ID	Average Distance
84	bonifico a vs.fav. saldo ft 7 del 04.04.25 cro: bonifico n. disposto	transfer, credit	transfer, charge	[2, 89, 204]	61.2%
202	liq. inter. ant.f competen	interests/charges	commissions	[12, 70, 11]	69.9%
603	liq. inter. ant.f competen	interests/charges	commissions	[12, 70, 11]	69.9%
704	bonifico a vostro favore bonifico sepa da banca popolare di puglia e basilica t	transfer, credit	commissions	[23, 37, 9]	65.0%
706	mav int.	external services	tax	[107, 105, 55]	55.6%
719	bonifico a vostro favore bonifico sepa da banca pop. alto adige per	transfer, credit	commissions	[23, 37, 9]	63.2%

TABLE 11: OPTICS - Error list without outliers.

nificantly, often over 50%, apart for 2 cases, (ID 202 and 472). The smaller average distance obtained with simple k-NN can be easily explained: there is wider number of points to select from, while, after sampling, the set of available points in the training set is reduced.

Evidence shows that classification errors occur more frequently when the average distance to the nearest neighbor is higher, hence it is possible to select a threshold marking as doubtful the descriptions with the average distance higher than that.

The effect of including outliers in the training set is exemplified by test ID 472 and 601 in Table 10, When outliers are included they are wrongly classified and their average distance to the nearest neighbors is less than half the value of standard errors. Once

outliers are excluded as in Table 11, these two test ID are correctly classified and the average distance to the nearest neighbors is everywhere above 55%, as the only available neighbors are those belonging to well-defined groups, thus significantly worsening their average distance.

This evidence suggests an important development: it would suffice to introduce a threshold for the acceptable NN average distance, to isolate classifications to be considered "doubtful" to be checked manually or to be confirmed with a rule-based approach.

Conclusions

To overcome the limitations of the rigid and time-consuming rule-based classification of bank transactions, this section presented

a Machine Learning approach, based on the K-Nearest Neighbors algorithm. The model relies on the normalized Levenshtein distance enhanced with text pre-processing techniques, such as converting the text to lowercase, eliminating irrelevant sequences of digits or text and truncating long descriptions.

With time, the size of the training set can increase significantly. To mitigate computational effort and scalability issues of the k-NN approach with large datasets, Hierarchical and OPTICS clustering methodologies were tested focusing on dataset reduction. After clustering transactions, representative points from each cluster were selected to form the new training set, reducing the original dataset size by 5-6 folds. The results obtained were very promising and can be summarized as follows:

1. The test performed using the full dataset, with "leave-one-out" method, achieved a surprising performance of 99% (6 transactions misclassified out of 834).
2. The test on the reduced dataset shows performances in line with those of the full dataset, also when excluding outliers.
3. The average distance of the tested descriptions to their NN can be used to provide a confidence measure for the assigned class and to define a threshold identifying doubtful classifications on which to concentrate direct controls.

In summary, the classification performances obtained with both clustering approaches are significative and consistent with those obtained using the entire training set, confirming that the methodology is scalable and reliable also with large training datasets. In addition, the average distance from the nearest neighbors, calculated for each test descriptions, provides a measure of the confidence of the assignment, sug-

gesting which of them should be submitted to a direct manual check.

Classifying Transaction with LLM

In addition to the previous classical methods, this chapter considers modern LLM techniques for transaction classification. Unlike classical methods, advanced Natural Language Processing (NLP) algorithms can capture contextual meaning instead of focusing on isolated words or symbols. This may be useful when applied to the description of financial transactions, where the order of keywords may be relevant for the classification.

The vector representations generated by using **contextual embedding methods** adapt dynamically to their relative meanings, based on the context where they are inserted. For example, the word "bank" is represented differently in "river bank" than in "money bank." This represents a substantial improvement over static embedding approaches, which assign a fixed vector to each word, regardless of its meaning, thereby failing to capture polysemy and contextual nuance.

A major architectural breakthrough occurred with Google introducing the **Transformer** model. Transformers eliminate the need for sequential token-by-token processing, enabling **parallel computation** across all input positions. The innovation point is the **self-attention mechanism**, which allows to directly model relationships between any pair of tokens, regardless of their distance in the text. [14] This improves both computational efficiency and its ability to represent global context. The Transformer architecture has become the foundation for nearly all state-of-the-art NLP systems, in particular after introducing the Universal Language Model Fine-tuning (ULMFiT), an effective learning method that can be applied to any NLP task, which significantly outperforms the state-of-the-art on

text classification tasks, reducing the error by 18-24%. [8] The Transformer architecture includes three variants: encoder-only, decoder-only, and encoder-decoder models, all built on the same attention-based foundation.

- *Encoder - models* are designed to improve text understanding and representation and are well-suited for tasks like classification;
- *Decoder - models* (e.g., GPT, developed by OpenAI) generate text autoregressively, making them ideal for content generation;
- *Encoder - decoder* models combine both components and are effective for sequence-to-sequence tasks such as translation and summarization.

This work is focused on the BERT model, particularly appropriate for classification task [6] ⁶.

BERT consists of many pre-trained transformer layers, and task-specific adaptation can be achieved by training the parameters of the final layer to a specific sample.

A challenge in applying modern NLP methods to this study lies in the limited number of training samples, approximately 850, insufficient to provide a robust training to an advanced deep learning model, with millions — or even billions — of parameters. While traditional string-matching techniques, such as the **Levenshtein distance combined with k-Nearest Neighbors (KNN)**, require less extensive training sets as they rely solely on character-level similarity, ignoring linguistic meaning. On the other hand, Transformer networks require extensive corpora to generalize test effectively, consequently it is necessary to adopt frameworks to enhance procedure to function well in low-data environment. In additions, bank descriptions are

typically short, loosely structured and often contain misspellings, abbreviations, or domain-specific jargon. Characteristics difficult to model effectively without substantial amounts of labeled data. For this reason no pretraining was performed and the raw BERT output was used instead. In addition, BERT was combined with a more traditional **Bag-of-Words (BoW)** representations. Pretrained language models already encode rich semantic information learned from large-scale corpora, enabling them to perform effectively even when fine-tuned on relatively small datasets. When coupled with simpler lexical representations, they provide a balanced framework capable of delivering reliable classification performance under data-scarce conditions.

The analysis is outlined in the following paragraphs by presenting:

- An outline of **text Tokenization**;
- The **spaCy environment**;
- The Bag-of-Words representation and **the BERT embeddings**, including a measure of **the confidence level** of classifications in the LLM environment;
- The test **results and conclusions**.

Tokenization

Tokenization is the process of segmenting a text string - such as a sentence or a paragraph into smaller, meaningful units called *tokens*. These units may correspond to complete words, word fragments (e.g., prefixes, suffixes), frequently recurring character sequences, individual characters or, less commonly, multi-word expressions. Serving as the fundamental units for computational models, tokens enable downstream machine learning systems to analyze and manipulate text in a structured manner [10]. To represent sentences as ordered sequences of tokens, early neural approaches

⁶BERT is developed by Google and widely adopted in the industry. It is a state-of-the-art language representation model trained on large multilingual corpora, including Italian, using self-supervised objectives such as masked language modeling and sentence relationship prediction.

introduced the use of encoders which are shown to learn a semantically and syntactically meaningful representation. Architectures such as **Recurrent Neural Networks (RNNs)** — including more advanced variants like **Long Short-Term Memory (LSTM)** networks and **Gated Recurrent Units (GRUs)** — process text sequentially, one token at a time by updating a **hidden state** that summarizes processed information [5].

This mechanism enables the model to capture word order and short-medium-range dependencies of tokens within a sentence, however, they tend to be slow to train and often struggle with very long-range dependencies due to issues such as vanishing gradients. Several established algorithms enable models to cope effectively with rare or unseen lexical items while maintaining semantic coherence. **WordPiece**, for instance, became prominent through its adoption in BERT as it provides a balanced trade-off between vocabulary size and linguistic flexibility [6].

Handling alphanumeric sequences, contained in transaction descriptions, may carry information relevant for classification. BERT handles such inputs through its sub-word tokenization process, which decomposes numeric and alphanumeric strings into smaller units and stores them in its vocabulary. These components are then embedded and contextualized, allowing the model to capture recurring structural patterns and incorporate them into the resulting embeddings. Through the construction of a consistent vocabulary — comprising all unique tokens appearing within a dataset — tokenization ensures standardized treatment of linguistic units across inputs.

In contrast, the Bag-of-Words approach, also adopted in the analysis, typically relies on standard tokenizers such as those provided by spaCy, which split text into whole words disregarding context. As a result, BERT captures semantic and contextual relationships, whereas Bag-of-Words represents text as simple word counts with

no contextual awareness.

One of BERT's core strengths is its ability to handle noisy and unstructured text inputs. By feeding resulting token sequences into ELMo and BERT, NPLs ability to recognize similarity between words, even with limited occurrences in the training data, is strongly improved, as contextual embeddings capture fine-grained language patterns. [11] Through its sub-word tokenization mechanism, BERT breaks unknown or distorted words into smaller units that become part of its learned vocabulary. This means that, even misspelled terminology or compound domain terms, can still meaningfully contribute to the final embedding. For this reason BERT is especially suitable to model bank transaction descriptions where numeric and mixed-format tokens appear frequently and carry domain-specific meaning, like the descriptions analyzed in this work.

The SpaCy Environment

One of the most established libraries for natural language processing is **spaCy**, which offers robust and linguistically informed tokenization, designed to efficiently split text into meaningful units. In our case spaCy is used to tokenize descriptions within the Bag-of-Words framework, since BERT uses its own algorithm (WordPiece) not suitable for the current purpose. In particular, spaCy provides a flexible framework enabling users to extend or customize tokenization rules.

SpaCy's tokenization process is primarily **rule-based** and can be enriched with pre-defined *special cases* associated with the selected language model, Italian in this case. These rules manage how the text is segmented into tokens while accounting for language-specific patterns, such as abbreviations, clitic attachments and punctuation conventions. It handles abbreviations through **language-specific tokenization rules** and **special-case exceptions**, rather than through a fixed list of

abbreviations. In the current work the following rules were applied:

A. Tokenization Rules and Exceptions

– SpaCy’s tokenizer includes predefined *special cases* that treat certain sub-strings, such as common abbreviations, as single tokens. For instance, “SRL” and “SPA” can be written in different ways like “S.p.A” or “S p A ”. Default tokenization algorithm may treat this as separate symbols (‘s’, ‘p’, ‘a’) and miss relevant information. To ensure consistent tokenization of these abbreviations, a special rule is added.

B. **Format Sensitivity** – the original word format (upper/ lower case letters) *is preserved* during tokenization and the generation of the vector embedding. This allows the model to distinguish between upper and lower-case formats of the same word, which can provide further meanings for the following classification tasks. It should be noted that in Levenshtein distance approach only lower-case words were used to avoid mismatching, whereas BERT embeddings captures this information.

The Bag-of-Words Representation and the BERT Embedding

The *Bag-of-Words* (BoW) model is a foundational method in natural language processing that represents a text document as a numerical vector based on the frequency of its words. [15] The model treats the text as a “bag” of individual terms, thereby disregarding grammar, syntax, and word order, assuming that, *“the meaning of a document can be inferred from word counts alone.”*

The construction of a BoW representation begins with the creation of a **vocabulary** derived from the entire corpus under analysis. This vocabulary consists of all *unique* tokens or, in some cases, sequences of tokens (n-grams), appearing across the dataset.

Each unique token or n-gram is then treated as a distinct **feature** in the resulting vector space. Consequently, every document is mapped onto a vector, where each element corresponds to the frequency with which the associated token occurs.

Because different inflected forms of a word (e.g., *work, working, worker*) may refer to the same underlying concept, many workflows apply **lemmatization** prior to frequency counting, a process that reduces words to their canonical form. As noted in computational linguistics, *“lemmatization consolidates morphological variants to ensure consistent representation of concepts”*. The spaCy library provides efficient and linguistically informed tools to perform such lemmatization. However lemmatization was not considered in this study because in transactions different words referring to the same concept may carry relevant information which should not be lost.

For each individual text segment (e.g., a product description), the BoW procedure computes the frequency of every vocabulary ‘word’ within that segment. The resulting vector is typically **sparse**, meaning that most values are zero, since any single text contains only a fraction of the full vocabulary. A sparse representation is a characteristic feature of BoW-based encoding and reflects the high dimensionality and lexical diversity of the selected text. The following paragraphs shows how to create advanced contextual embeddings that can capture semantic and contextual information.

The BERT Embedding

To capture the semantic and contextual information embedded in bank-transaction descriptions, this study employs a pre-trained BERT Transformer exclusively as an **embedding generator**, transforming each sparse transaction description into a dense, fixed-length vector appearing as a rich and informative feature representation.

The model’s capacity to infer contextual meaning can be illustrated using the transaction description: “ACCR. BONIFICO EU

COD. DISP.: 0124".

When processed by BERT, the output is a single dense vector that captures the latent relationships among the abbreviated and domain-specific terms. For instance, "ACCR." is a shortened form of *accredito*, while "BONIFICO" combined with the token "EU" suggests an association with *EU-area SEPA transfers*. The resulting embedding occupies a region of the vector space proximal to other transaction descriptions expressing semantically analogous operations, related to incoming SEPA transfers, such as:

- "accredito bonifico SEPA";
- "transfer received EU";
- "movimento bancario in entrata";
- "incoming payment code 1234".

Sentence classification is the most direct and intuitive applications of BERT, as illustrated by Jay Alamar in [1] In practice this simple embedding and fine tuning, leverages BERT's pre-trained linguistic knowledge and often yields strong performance with minimal specific engineering.

Enhanced Bag-of-Words Representation

Traditional approaches like Term Frequency-Inverse Document Frequency (TF-IDF) and Bag-of-Words, offer simple and computationally efficient means of representing term importance within a document, but they lack the ability to capture semantic relationships. Recent works have shown that these lexical representations can be enhanced through the integration of pretrained embeddings capable of modeling semantic and contextual information as in [4] where the BoW representation and BERT are mixed through a weighted feature-level strategy. Empirical results demonstrate that hybrid models achieve superior training performance and attains the highest test accuracy among the evaluated configurations.

In practice, in this study:

- The text was tokenized using spaCy, to ensure that all tokens are preserved, including punctuation, domain-specific terms, and alphanumeric strings;
- These were then converted into a BoW vector reflecting the raw frequency of each token in the description (i.e. without applying TF-IDF weighting or normalization);
- Similarly, transaction descriptions were also encoded using BERT to generate a deep semantic embedding;
- Finally, the two vectors were concatenated to create a combined feature vector leveraging both high-level contextual understanding and straightforward term-frequency information.

By leveraging both contextual and lexical cues, the hybrid model provides a more comprehensive and discriminative representation of textual content. This joint vectors represent the inputs for the classification task.

The selection of the classification models depends on several factors, in particular: data availability, separability of classes and a confidence interval of classifications.

- **Data availability** is a key driver for model selection. Given the relatively small size of the available sample, classical Machine Learning classification models are considered as they perform well both on small and large dataset. However, the framework is built to accommodate large amounts of data which allow to consider more advanced methods, particularly BERT embeddings for domain specific classification tasks.
- **The confidence level of classifications** requires to define a metric identifying poorly classified descriptions. Classifications with a low confidence score should be investigated manually.

Potentially, with a larger sample of data, ensemble framework could be used, i.e. run multiple models and average their predictions.

Measure the Confidence Level of Classifications

With the K-NN algorithm to identify a measure for the confidence level was straightforward and intuitive: the average Levenshtein distance to the closest neighbors, representing the degree of overlapping of the test description with the selected sentences. In the LLM framework however, the output of the classifications process consists of the a probability distribution Y of the different tags being assigned to the test description through a multinomial logistic regression;

$$Y = \begin{bmatrix} P(tag_1|x) \\ P(tag_2|x) \\ \dots \\ P(tag_n|x) \end{bmatrix} \quad \text{with} \quad \sum_i P(tag_i) = 1,$$

with the tag effectively assigned to the test description being that with the highest probability,

$$tag(x) = Y_i \quad \text{where} \quad P_i = \max\{P(tag_i|x)\}.$$

In this case the probability vector Y is given by the **Softmax** function 1, one of many functions mapping non-normalized scores onto probability vectors, but one particularly useful in IO models, where only the right prediction is relevant and not the transparency or accuracy with which the output is modeled.

$$Y_i(x) = \frac{\exp^{-a_i x_i}}{\sum \exp^{-a_i x_i}}. \quad (1)$$

Clearly, the confidence level of this assignment depends not only on its probability value, but also on how the remaining probability is spread across the other classes. One possible approach to measure uncertainty in the LLM world, is to look at the **predictive entropy** equation 2, representing the

average amount of information within the predictive probability distribution:

$$E = - \sum_{i=1}^n p(tag_i) \log(p(tag_i)), \quad (2)$$

where the summation is over all possible tags that Y can assume. The predictive entropy reaches its lowest value when all classes are associated with equal probability, indicating complete uncertainty. Conversely, in case of a very confident classification, it reaches its highest value of near-0, when one class is predicted with near certainty (a probability of near-1).

Test Results and Conclusions

Cross Validation

To obtain reliable evaluation results, cross-validation is applied by training and testing the classification model on multiple data splits. Leave-One-Out Cross-Validation (LOOCV) can be effective for small datasets and lightweight models as was described in Chapter 1. However, it becomes computationally impractical for more complex architectures with advance classification models. Here, instead, we use Stratified Cross Validation.

In **stratified cross-validation**, the dataset is partitioned preserving the relative proportion of each key variable between the training and testing folds. This means that each partition reflects the original class distribution, an essential requirement when dealing with imbalanced datasets or rare categories. The stratification constraint ensures that each fold is a representative sample of the full dataset with respect to the target variable. This reduces variance in performance estimates and avoids biased evaluations that can occur when certain classes are over or under-represented in the test sets. By maintaining consistent class proportions across folds, stratified cross-validation produces more stable and meaningful comparisons between different models.

The model is then trained and evaluated multiple times across different stratified folds, yielding a distribution of performance metrics, such as errors, accuracy and confidence level. This richer set of results increases statistical robustness and enables the detection of systematic mis-classification patterns recurring across folds. Categories distribution is not particularly concentrated in one or two classes. Moreover, errors in classification are usually encountered among the most frequent categories: "bonifico addebito and commissions". It implies that current model allows to separate more specific categories of transactions. Consequently, stratified cross-validation provides a more comprehensive and reliable assessment of a model's generalization capabilities than either a single train-test split or non-stratified cross-validation methods. Its effectiveness is particularly pronounced when working with datasets characterized by class imbalance, limited size, or heterogeneous label distributions.

Results and Conclusions

This research provides a comprehensive evaluation of automated systems for classifying bank transactions, a tedious but vital component in the reconciliation of invoices and financial records. The study proves that the integration of AI and ML techniques offers a precise, consistent, and scalable solution for financial control. Both traditional distance-based methods and modern contextual models can achieve a significant accuracy (99% and above), comparable to manual rule-based systems, along three main dimensions. It should be noted that the model performs well even for transaction categories with a small number of training samples; therefore, the accuracy metric is considered a reliable indicator of overall performance.

- **Scalability and Efficiency:** the first part of the work demonstrates that the **K-Nearest Neighbors (k-NN)** approach, using Levenshtein distance,

is highly effective for short text segments. Furthermore, the application of **HAC and OPTICS clustering** successfully help reducing the training dataset size by **5 to 6 times** without compromising performance, providing a clear path for managing scalability as transaction volume grows.

- **Semantic Depth:** the second part reveals that **Large Language Models (LLMs)** like BERT can interpret the contextual meaning and domain-specific jargon of financial data, even when training samples are limited. The results from the LLM-based approach, reported in Table 12, demonstrate that integrating BERT embeddings with Bag-of-Words (BoW) representations successfully overcomes the limitations of small, unstructured training datasets in the financial contexts. By combining high-level contextual understanding with straightforward term-frequency information, the hybrid model achieved an average accuracy exceeding 99% across eleven iterations of stratified cross-validation.
- **Practical Implementation:** a significant outcome for both methodologies is the ability to generate **confidence metrics**, which allow organizations to focus manual efforts exclusively on "doubtful" cases. **Predictive entropy** was selected as a measure of the confidence level for classifications in the LLM environment. A key finding of this work, was the effectiveness of the 70% confidence threshold, above which zero classifications errors were found, hence serving as a reliable indicator for triggering manual controls for the few classifications obtaining a smaller confidence level.

This confirms that advanced NLP techniques can be adopted also for low-data environments to capture the nuanced meaning of bank descriptions, ensuring a robust

Split	Above Threshold		Below Threshold		Whole Test Set			
	Points	Errors	Points	Errors	Accuracy	Accuracy	Correct Tests	Wrong Tests
1	194	0	15	3	80%	99%	0.91	0.47
2	198	0	11	1	91%	100%	0.91	0.66
3	197	0	12	1	92%	100%	0.91	0.63
4	197	0	12	0	100%	100%	0.91	-
5	199	0	10	1	90%	100%	0.92	0.66
6	200	0	9	1	89%	100%	0.91	0.58
7	195	0	14	0	100%	100%	0.90	-
8	191	0	18	8	56%	96%	0.91	0.39
9	201	0	8	2	75%	99%	0.92	0.40
10	196	0	13	1	92%	100%	0.91	0.69
11	199	0	10	1	90%	100%	0.91	0.22

TABLE 12: LLM results by iteration.

and trustworthy automated classification process.

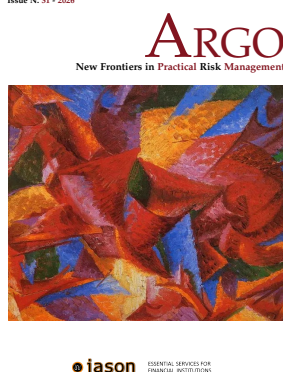


References

- [1] **Alammar, J.** *The Illustrated BERT, ELMo, and co. (How NLP Cracked Transfer Learning)*, Hands-On Large Language Models, 2018.
- [2] **Almar, J. and Garcia, J.** *The Impact of Artificial Intelligence and Machine Learning on Management Accounting*. International Journal of Multidisciplinary Research and Publications, ISSN 2581-6187, pp. 114-121, 2025.
- [3] **Ankerst, M. et al.** *Optics: ordering points to identify the clustering structure*. ACM SIGMOD Record, Vol. 28, issue 2, pp. 49-60, 1999.
- [4] **Aprilio, P. et al.** *Hybrid Feature Combination of TF-IDF and BERT for Enhanced Information Retrieval Accuracy*. Jurnal Informatika dan Sains, June 2025.
- [5] **Cho, K. et al.** *Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation*. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 2014.
- [6] **Devlin, J. et al.** *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. October 2018
- [7] **Fieberg, C. et al.** *Machine Learning in Accounting Research*. Springer, pp. 105-124, September 2022.
- [8] **Howard, J. and Ruder, S.** *Universal Language Model Fine-tuning for Text Classification*. January 2018.
- [9] **Munoz, J., Jalili M. and Tafakori, L.** *Hierarchical classification for account code suggestion*. Knowledge-Based Systems, Volume 251, September 2022.
- [10] **Phan, C. and Singh, I.** *Understanding WordPiece Tokenization: An Approach to Subword Units*. Collection of Machine Learning Interview Preparation materials, 2026.
- [11] **Salunke, A.** *Enhancing Contextual Understanding in NLP: A Subword Tokenization Approach with ELMo and BERT*. June 2023.
- [12] **Sebastiani, P. and Perls, T.** *Detection of significant groups in hierarchical clustering by resampling*. Frontiers in genetics, Vol. 7, pp. 144, 2016.
- [13] **Tootian, P.** *The Impact of Artificial Intelligence Application on the Accounting Profession*. 2023
- [14] **Vaswani, A. and et al.** *Attention Is All You Need*. June 2017.
- [15] **Zhang, Y. et al.** *Understanding bag-of-words model: a statistical framework*. International Journal of Machine Learning and Cybernetics, August 2010.

In the previous issue

Issue N. 31 - 2026



Issue N. 31 - 2026

CREDIT RISK

An Interpretable Machine Learning Framework for Credit Risk Satellite Models

Revenue Based Finance: an Evaluation Framework

INSURANCE RISK

Multilevel Monte Carlo for Solvency II SCR: Efficient Nested Simulation

ENVIRONMENTAL, SOCIAL AND GOVERNANCE RISKS

SFDR 2.0 and EU Taxonomy in Asset Management Field

Last issues are available at www.iasonltd.com/research/argo/