The Case for Integrating Self-Organizing Maps with Large Language Models for Insurance Analytics

Ivelin Zvezdov, Head of Aggregate and Financial Modeling, Verisk CRS, April 2026

Introduction

Machine learning and artificial intelligence applications have already made a significant mark in quantitative and catastrophe risk modeling for (re)insurance analytics. Genetic algorithms perform portfolio optimization objectives. Self-Organizing Maps perform tasks for detecting erroneous data and filling sparse information matrices. The arrival of Large Language Models trained with the same algorithms creates options for integration. Optimal balance in leverage of the private machine learning systems and generic vendor large language models is in the best interest of market practitioners.

1. The Business Case

Clear trends in machine learning and artificial intelligence are converging in a growing reinsurance industry. This process needs attention and reconciliation. For three decades, specialists in insurance and finance have built machine learning systems to solve various complex problems. Prominent and recognizable feats include the wide implementation of Genetic Algorithms for portfolio optimization. Less well known is the application of Self-Organizing Maps [SOM]. The latter are highly capable of consuming unstructured, multi-dimensional data for the purpose of classifying and ordering it by properties derived from the key attributes of these large deposits of information. While performing this task, SOM reduces dimensionality, installs order, and learns in the process. These maps are neural networks by definition and are capable of unsupervised learning and self-correction. The Finnish computer scientist and mathematician Teuvo Kohonen pioneered the algorithms in the late 1980s and early 1990s. This proliferation of private machine learning systems is our first and well-established trend.

Enter Large Language Models [LLM], built and delivered by Big Tech vendors. These have all the capabilities of neural networks and genetic algorithms. However, the advantages of proprietary machine learning systems, trained and refined over time, are manifold. Firms have assessed and proven these internal and private systems over years, and by now they require minimum supervision from practitioners. Users have straightened the errors and polished up performance through countless hours of training and production. Secondly and more significantly, these systems contain the topology of risk factors of the firm. This is the core business model and philosophy, which firms protect keenly. Hence the solution of coexistence between private machine learning systems and vendor Large Language Models is integration. This is our second newly exposed trend.

Lastly but not least of all, we have the expansion of the reinsurance business into developing and growing markets and regions. This is our third, dynamic and well-recognized trend. We will take a case in point with the oldest reinsurance contract, the Quota Share of Catastrophe Loss.

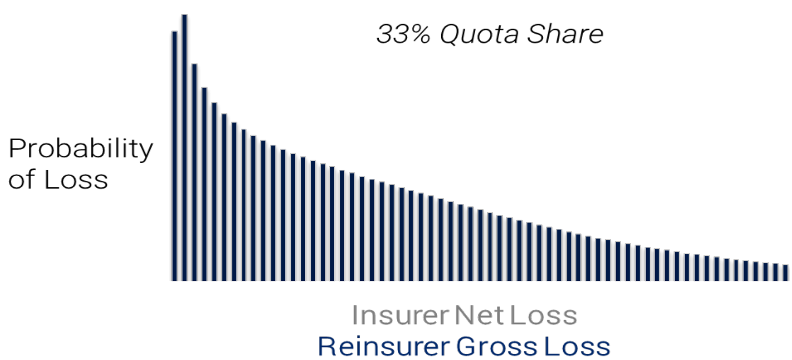

Figure 1, A 33% Quota Share treaty applied 'from the 1st dollar' on the insurer distribution of gross loss, resulting in 33% of ceded loss to the reinsurer, and 67% net retained.

The treaty is a fitting instrument to minimize earnings volatility, while supporting ambitious market share targets. This has been the case since the time when Venetian and Genovese bankers reinsured Mediterranean and Black Sea trade. From then to now volatility is particularly important in a growing market where underwriting targets keep up with fast expansion and a healthy degree of uncertainty. From then to now, reinsurance has always been an information business. The quality of estimates in exposure-at-risk by a process of quantifying the amount of risk a cedant carries and how much of it a reinsurer assumes determines the accuracy of pricing, reserving, and capital allocation. A large share of consequential information that drives exposure lives in unstructured data formats: government circulars, regulatory filings, rating agency reviews, accounting standards, broker advisories, and increasingly, satellite physical damage assessments.

Two prescient cases in point are Malaysia and Indonesia with 5-to-8 percent annual growth in Gross Underwritten Premium. This makes the region a dynamic and demanding marketplace.

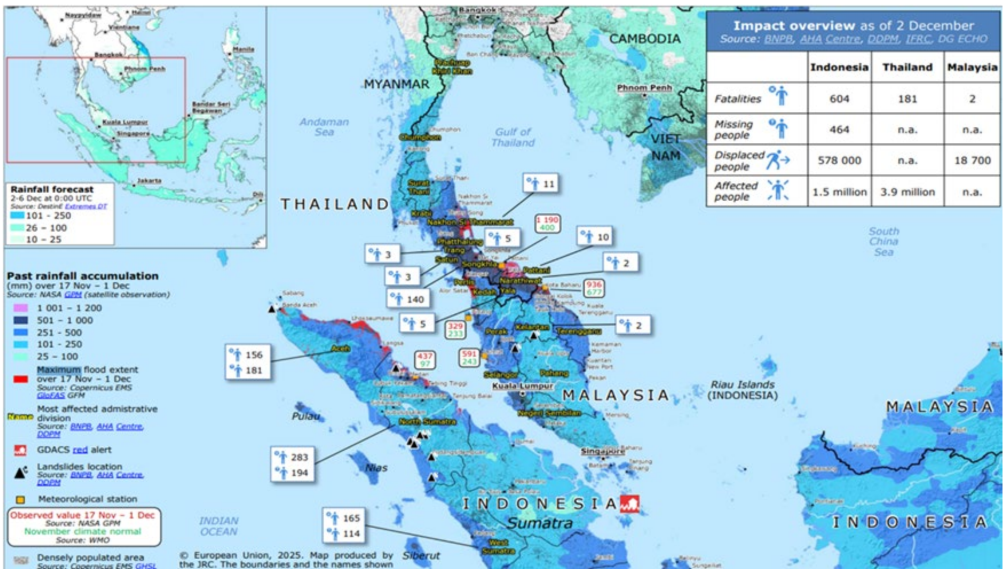

Figure 2, Impact of flooding in Malaysia and Indonesia during November of 2025. Exhibit produced by reliefweb.

Reinsurance cycles move quickly. In an environment of sparse data and limited historical experience simplicity of structure and instantaneous transparency of pricing techniques become an advantage. Quota Share is the reinsurance contract with lowest operational cost and clearest, most stable and most recognizable price. It reduces volatility across the entire book and the entire risk tail of the business. Under the linear and proportional premium-making and loss-ceding rules of the treaty, reducing uncertainty and error in underlying exposure, directly and surely reduces uncertainty in loss outcomes and in earnings volatility.

The Self-Organizing Map for data processing is a tried and tested algorithm which can streamline the validation of the ceded exposure of the insurer to its partner the reinsurer. It reduces multi-dimensional data to two-dimensional surfaces by pre-selected rules, while learning, training and self-correcting. This makes it perfect for ingesting substantial amounts of exposure and premium records, historical loss and claims, rates, and indices well in synchronicity with qualitative data and narrative from brokers, government, and accounting agencies. SOM is connectable directly to exposure databases and lakes. Self-organizing and self-learning layers process volumes of ingested data to create an exposure map of linear variables of business interest.

In this case the aim is to vet, correct and fill in sparse data on exposure variables of key business concern such as insurable values, deductible amounts, and spatial coordinates of risks. Then the system overlays the output from SOM onto the ceded exposure of the insurer and the procedure itself effectively executes validation, correction, and self-adjustment. As a result, the integrated system reduces uncertainty and error in the targeted data repositories of insurable exposure. Through the proportional nature of the Quota Share contract this has an immediate multiplier effect on containing uncertainty and volatility in earnings.

2. The Integrated System

There are various integration concepts capable of addressing and reconciling the intersection of these three trends described so far. We continue by reviewing in light non-technical terms one such concept in the form of a three-layered system. Clear and optimal understanding of architecture allows to partition tasks, components, and layers across this multi-dimensional system.

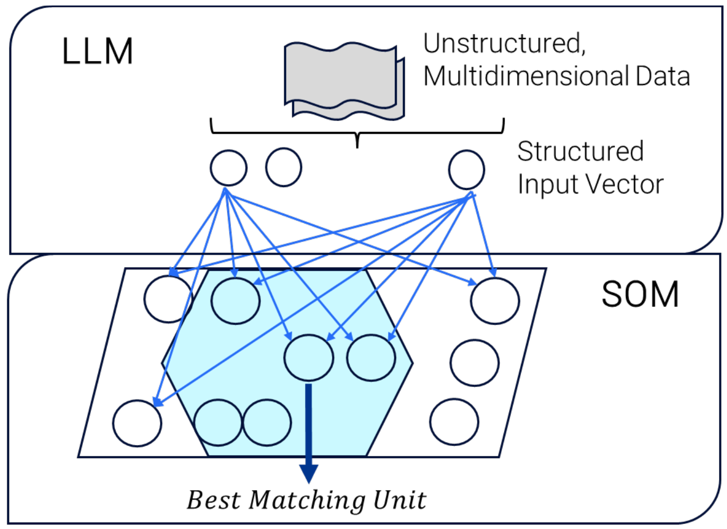

Figure 3, A generalized view of the integrated system

Layer	Component	Function
Layer Zero	Data Sources	Government documents, rating agencies data, accounting standards, broker reports, satellite feeds
Layer One	Vendor LLM	Ingest documents; extracts vectors of structured signals for learning and modification of business variables
Layer Two	Private SOM	Maintain risk factor topology; classify signals and apply them; update variable-of-business interest
Layer Three	Feedback Loop	Compare realized variables to modified; recalibrate all parameters over time

Table 1, Three-layer integration architecture, from raw data sources to closed-loop learning.

Layer Zero contains and encompasses all unstructured, multi-dimensional data such as government agency reports and regulatory circulars, rating agency publications, accounting standards, broker summaries, and space satellite feeds.

In **Layer One** a vendor LLM consumes unstructured, multidimensional numerical and qualitative data. Its first task is to represent all data sources as vectors and surfaces $(x_1, \dots, x_n)$. Then to move and construct feature vectors X_i from these vectors. The system architect and expert user assign to each source a credibility weight λ which reflects the level of authority and reliability of the source.

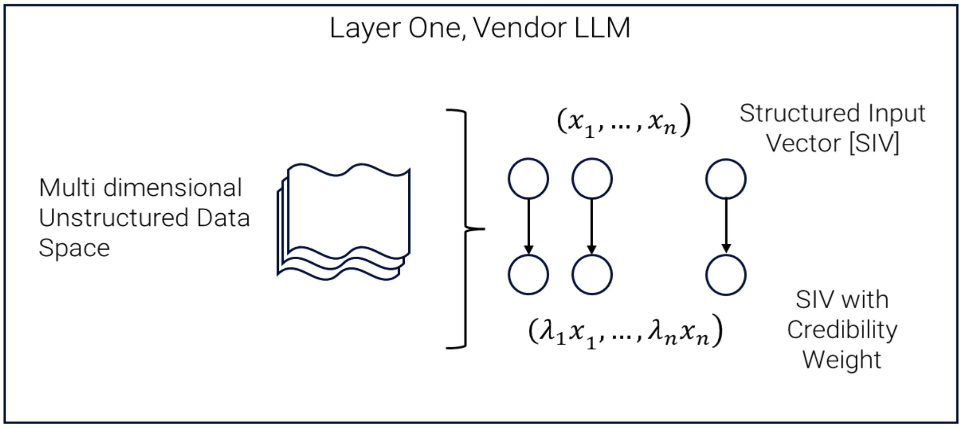

Figure 4, In Layer One LLM builds structured input data vectors with credibility weights from multi-dimensional unstructured data space.

When multiple sources $(x_1, \ldots, x_q | x_n)$ point to the same risk class or exposure variable-of-interest within the same processing cycle, credibility weighted average aggregates their signals, *equation 1*. By this approach, no single low-credibility data point can dominate the combined vector signal X_t.

$$X_{i,t} = \sum_{x=1}^{n} (\lambda_i * x_i) \tag{1}$$

The expert user will define a theoretical hierarchy with credibility weights and then revise it empirically through a feedback loop. This we describe further below. A decay function, where k is a user-defined positive time decay constant and t is time allows the architect to control the temporal evolution of credibility and trust assigned to the ingested data at time t_0.

$$\lambda_{i,t} = \lambda_{i,0} * exp^{-kt} \tag{2}$$

In this distribution of labor, LLM works out the feature vector X_t which maps to the SOM neural nodes in **Layer 2**. The trust level scaler λ_t above marks to every document and every unstructured data source $(x_1, \ldots, x_n)$ from which LLM constructs feature vectors. In this proposed concept the system architects and practitioners assign this trust variable, and the human user keeps control of a critical risk control and mitigation lever. The LLM role is strictly that of a translator. It converts rich, and potentially ambiguous, context interpreted language of regulatory documents and broker submissions into structured numerical vectors X_t. SOM in Layer 2 recognizes these vectors X_t as its input data feed.

In **Layer 2** proprietary SOM developed in-house by the firm retains control of definition and mapping for all risk factors and all business variables. It is essential that the business owner keeps the SOM grid proprietary. This is the practitioner's market-making guideline and philosophy of risk topology. We do not want to outsource this to LLM. This is the business model of the firm.

A prototype surface $y_0 = y_{0,1}, \dots, y_{n,q}$ is set up with data points sampled at random. This surface is our concept of the desired, final, and optimal state of variables of interest.

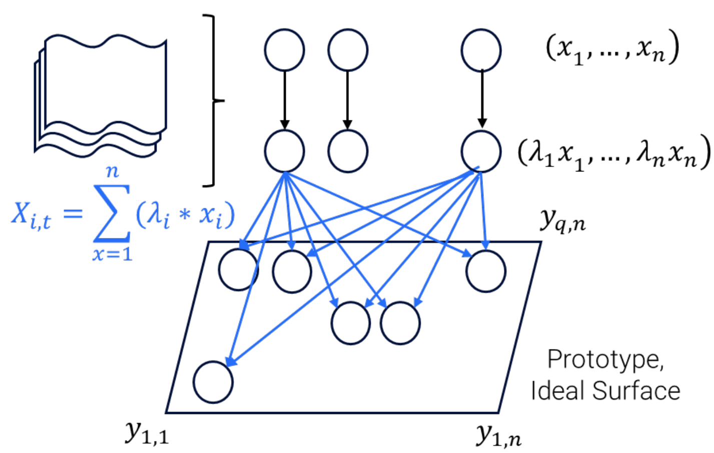

Figure 5, Structured Vectors $X_{i,t}$ map to prototype, ideal surface. This task occurs in the intersections of LLM and SOM. LLM can encompass it after few successful training cycles.

At time t initializing the first cycle of the system, a random sample function picks up data point x_i from the structured vectors X_t. The Euclidian distance between the current state x_i and the desired ideal state y_q becomes the Best Matching Unit [BMU] of the integrated system.

$$BMU_i = argmin\sqrt{\left(x_i - y_{n,q}\right)^2} \tag{3}$$

Through a purposefully designed neighborhood function h_t each BMU_i is recalibrated across all units $\{(i, \dots, q) | (x_i, \dots y_q)\}$ in vectors X_t and optimal surface $y_{i,q}$, with squared distance between map units (k, q), $dist^2(k, q)$.

$$BMU_{i+1} = BMU_i * \rho_{i,i+1} * exp\left(-\frac{dist^2(k,q)}{2\sigma^2(t)}\right) \tag{4}$$

$$where, h_t = exp\left(-\frac{dist^2(k,q)}{2\sigma^2(t)}\right)$$

The temporal variance $\sigma^2(t)$ plays control role to reduce or expand the connectivity of each BMU_i with the same units across the neighborhood map units. Lastly, we have an optional inter BMU correlation factor $\rho_{i,i+1}$. This can be a single, fixed factor for each process run or a detailed matrix with correlation weights for each-to-each neighborhood surface relationship. It is preferable that the system architect and expert practitioner define $\rho_{i,i+1}$, so that we introduce a control lever for the human user.

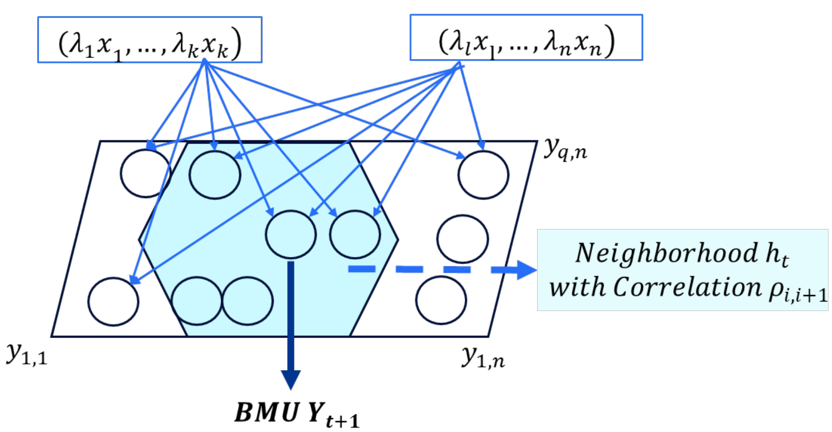

Figure 6, SOM selects Best Matching Unit [BMU] within the prototype surface $(y_{1,1}, \dots, y_{n,q})$ *using a neighborhood function and expert-architect defined correlation coefficient.*

This is how we come to putting together all the pieces and components of the core equation of proprietary SOM in detailed form, with a retrospective learning function $\varphi(t)$.

$$Y_{t+1} = y_{t=0} + \varphi(t) * \ h_t * \rho_{i,i+1} * argmin\sqrt{\left(x_i - y_q\right)^2} \tag{5}$$

In simplified form using all components which we developed above we have.

$$Y_{t+1} = y_{t=0} + \varphi(t) * BMU_{i+1} \tag{6}$$

The core equations of SOM modify our exposure variable of interest. It is fully transformable in the context of (re)insurance industry practices.

Term	Context in (re)insurance
y(t=0)	Current variable-of-interest for **Layer 2** in ideal, initial state before arrival of signals from the Best Matching Unit.
Y(t+1)	Updated variable after SOM processes all signals from BMU with neighborhood connectivity, correlations, and retrospective learning.
h_t	A neighborhood function which acts to build the connectivity and surface structure among various signals pointing to the same variable of interest
BMU	This is a neural node whose profile optimally matches the incoming signal from LLM structured data vectors to its initial, ideal state.
$(x_1, \dots, x_n)$	Modification and signal vectors extracted by LLM from unstructured data, including satellite feeds, and prepared for consumption by SOM
$\rho_{i,i+1}$	The relationship of various versions, iterations, institutional or broker views of the same business metric or variable.

Table 2: Terms and definitions in the SOM core equation defined in the context of (re)insurance industry practices.

In **Layer 3** the process becomes machine learning through a feedback loop (retrospective) function $\varphi(t)$. It distributes corrections to correlated neural nodes and layers. The feedback loop function $\varphi(t)$ at run-cycle drives the recalibration channel. When new surface data becomes available at some future time cycle $t + i$, it corrects the SOM prototype surface $y_0 = y_{0,1}, \dots, y_{0,q}$ by calibrating to an actual observation $y_{t+i,1}, \dots, y_{t+i,q}$ through the neighborhood function h_t.

$$\Delta\varphi(t) = h_t[(y_{t+i,1}, \dots, y_{t+i,q}) - (y_{0,1}, \dots, y_{0,q})] \tag{7}$$

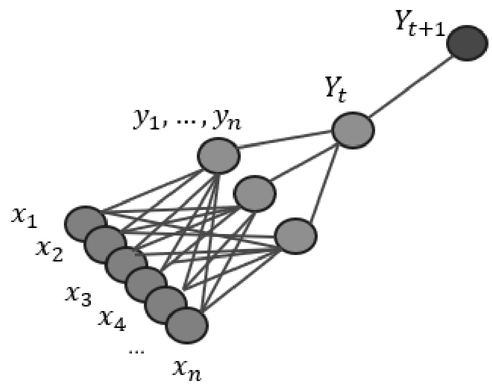

Figure 7, Generalized flow from unstructured data vectors $(x_1, \dots, x_n)$ *to learning layers* $(y_{t+i,1}, \dots, y_{t+i,q})$ *to Best Matching Unit* Y_{t+1}

One particularly powerful enhancement to the machine learning **Layer 3** is Retrieval-Augmented Generation. This process allows LLM to query the full ensemble of previously processed data whenever a new signal is ambiguous. The procedure examines the interpretation of similar data points in the past and what their impact was in the construction of feature vectors and surfaces.

This is as much as we will go into the mechanics and mathematics of SOM algorithms. There is a big and thriving literature and publications on the topic. For our purposes this is sufficient to distribute the main layers, components, and tasks of this concept system.

Lastly the modified variable of interest Y(t+1) propagates to a catastrophe modeling system, such as Verisk Synergy Studio where it enters the (re)insurance loss module and estimates reinsurer treaty and insurer retained loss and all probabilistic metrics such as Value-at-Risk with confidence α and portfolio exceedance distributions F^{-1}.

$$Treaty\ gross = ceded\ \% * \sum Y_{t+1} * exosure * damage \tag{7}$$

$$Insurer\ Net = \max\left(Insurer\ gross - Treaty\ gross, 0\right)$$

$$VaR(Insurer\ Net) = F^{-1}(1-\alpha)$$

Because the Quota Share contract is a linear and proportional reinsurance treaty, every improvement in the accuracy of our variable of interest Y_{t+1} translates directly and immediately into a more accurate loss estimate. There is no non-linearity to dilute gains. Improvements compound through the treaty structure and surface as reduced volatility in reported earnings. The Quota Share contract precisely delivers this outcome for rapidly growing markets such as these of Malaysia and Indonesia require.

3. Value added for the Market Practitioner

This proposed concept of integration architecture allows for comprehensive collaboration between human judgment and machine intelligence. The vendor Large Language Models are powerful but designed through generic training and calibration. They do not know anything specific about the risk topology and philosophy of the (re)insurance firm. The proprietary Self-Organizing Map fills this gap, designed and trained with the firm's own data, refined through the firm's own loss experience, and corrected through the firm's own feedback loop. The Large Language Model serves to support the Self-Organizing Map. The Self-Organizing Map supports the practitioner. And the practitioner through the assignment of trust and credibility variable and neighborhood correlation coefficient maintains oversight of feedback corrections and retains meaningful control over the entire system.

With this concept of integration, we have preserved the utility of business intelligence developed and refined in the firm in the form of a private machine learning system. We have coupled and integrated this with the new power and capabilities of Large Language Models. The outcome is a system which ensures that every material signal, regulatory, actuarial, and physical counts for impact in the exposure-at-risk model.

References

Hainaut, Donatien, 2018, A Self-Organizing Predictive Map for non-Life Insurance, European Actuarial Journal, December 2018

Kohonen, Teuvo, 2001, Self Organizing Maps, Springer Series in Information Sciences

Kohonen, Teuvo, 2013, Essentials of Self Organizing Maps, Elsevier Journal, Neural Networks (37), 2013

Wehrens, Ron & Kruisselbrink, Johannes, 2018, Flexible Self Organizing Maps in kohonen 3.0, Journal of Statistical Software, November 2018