Harnessing AI for Insurance: Real-World Success Stories

Paul Marron & Alexandre Boumezoued

FEBRUARY 13, 2025

Milliman

Solutions for a world at risk™

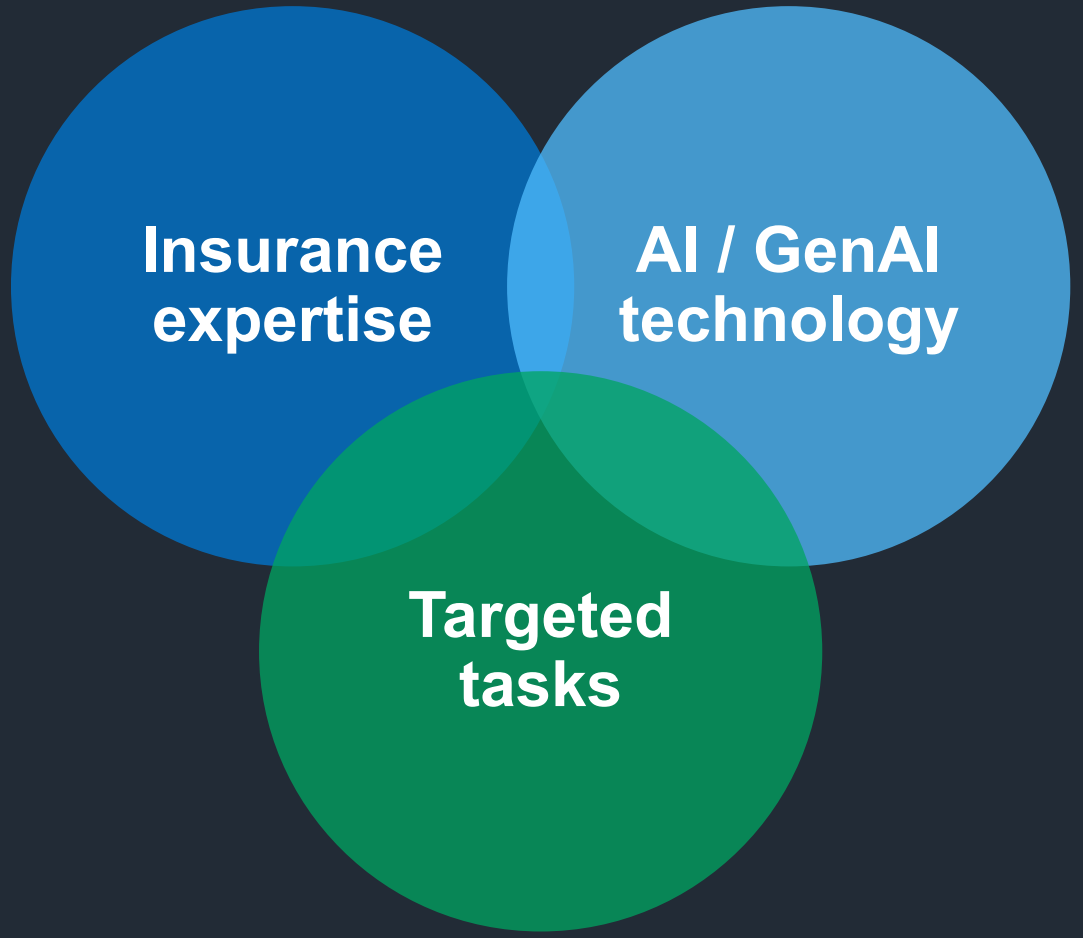
Harnessing AI for Insurance: Real-World Success Stories

- Main finding
- Five case studies
- The Gen AI checklist

Main Finding

Milliman

Value appears when expertise is combined with dedicated technological layers, applied to targeted tasks



Case Studies

Milliman

Large Language Models?

“Our general task-agnostic model outperforms discriminatively trained models that use architectures specifically crafted for each task”

— Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving Language Understanding by Generative Pre-Training. OpenAI.
This paper by OpenAI introduced the concept of Generative Pre-Training (GPT) for natural language processing tasks.

Case Study #1

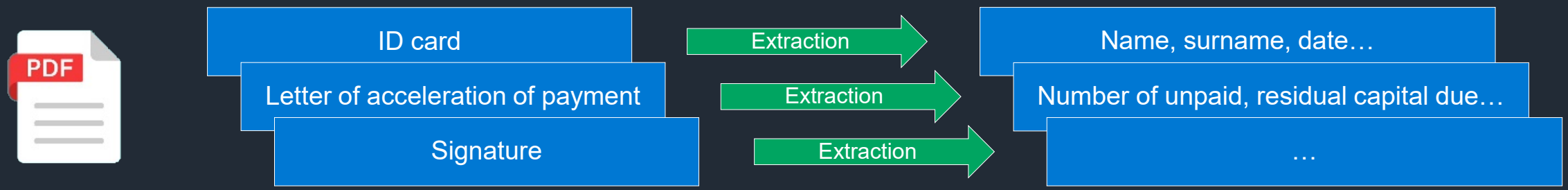
Automating Claims Processing

Milliman

Based on the knowledge of the process and its constraints, we identified a list of tasks that could be automated for increased efficiency and reliability

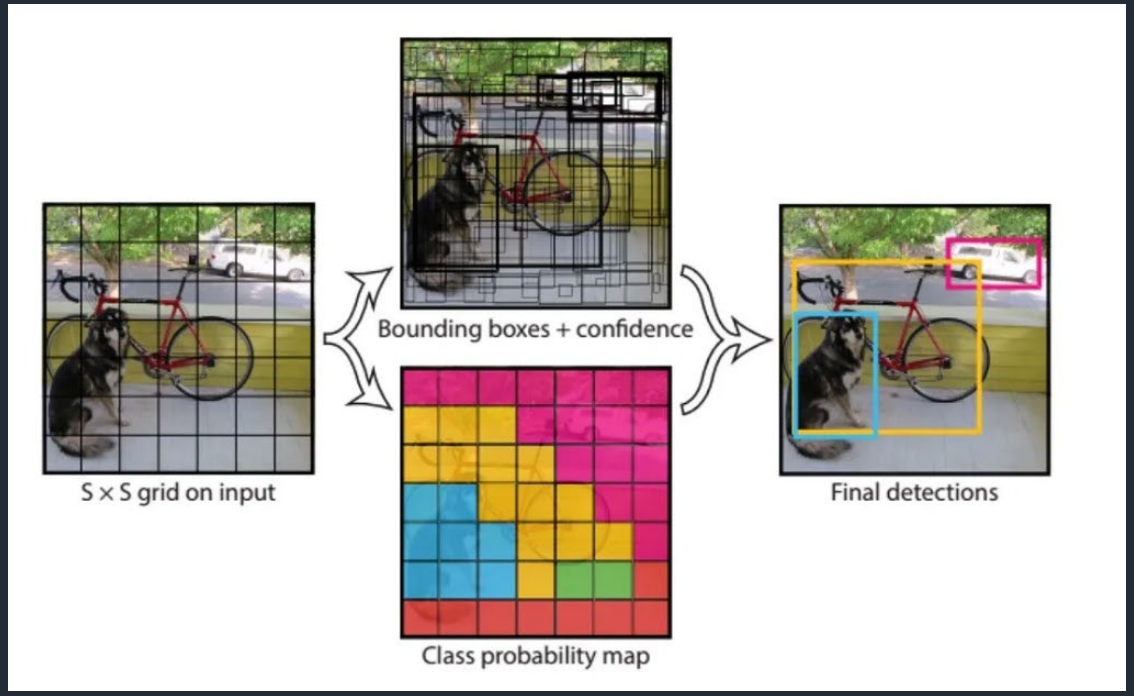
Creditor insurer: the decision-making process for compensation results from a thorough review of the claim file submitted by the bank, which consists of **numerous PDF documents**

Automating claims processing

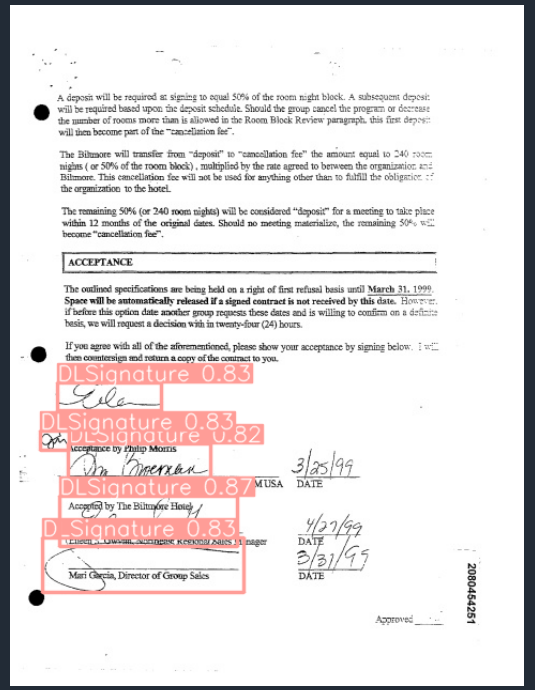


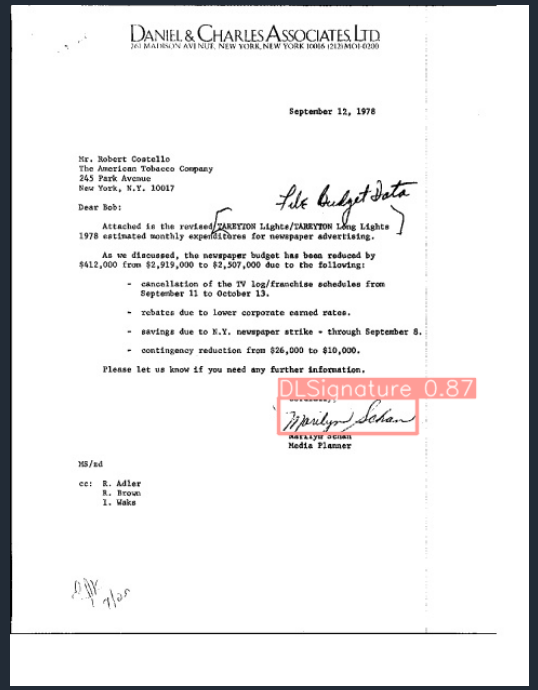
Solving the automation problem by training an open-source model to specific tasks, as opposed to using a general “off the shelf” model

Open source model dedicated to object detection



Example of signature detection after training





Extracting ID card information required a distinctive combination of technological layers

INPUT

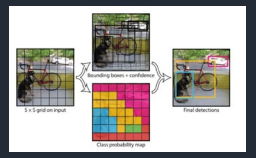
n-page PDF file

Conversion to images

Image list

Object detection algorithm

List of front and back pages of ID cards



Optical Character Recognition model

OUTPUT

Extracted information

- Nom : DURAND
- Prénom : ANTOINE PAUL HENRI
- Date de naissance : 01/01/1960
- Date de délivrance : 01/01/2012
- Date d'expiration : 01/01/2022

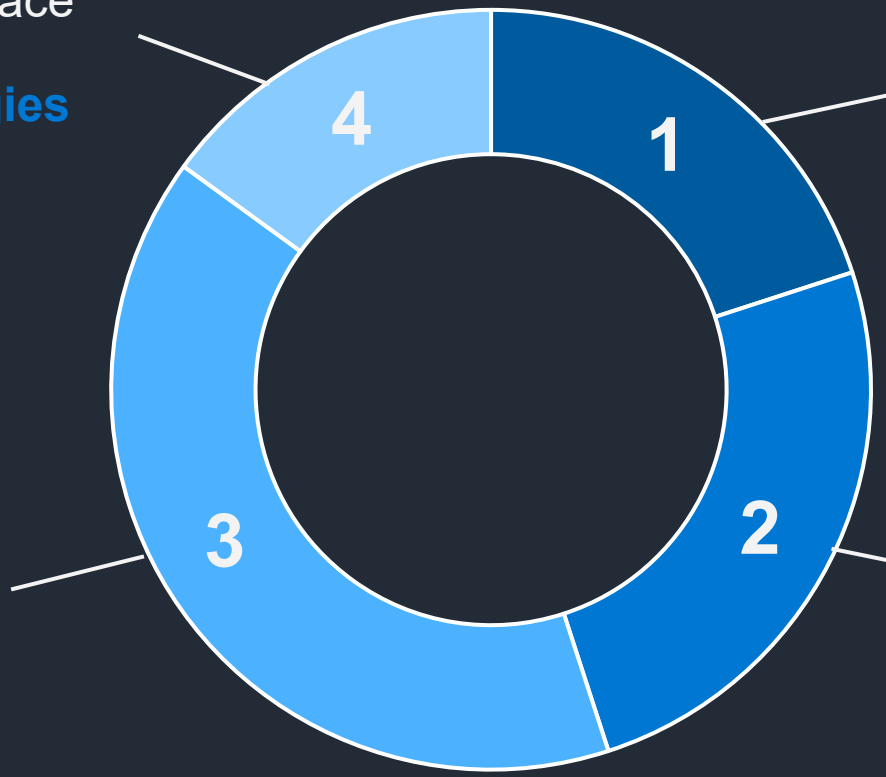
RegEx processing

String list

While algorithms deserve dedicated efforts, other workstreams are essential to project success: initial diagnosis, solution design, and change management

Effort Allocation for Project Success

Change management to embrace improved processes and the **ambition to reuse technologies in other areas**



Cartography & effective diagnosis of priority areas for automation is based on a **deep understanding of the current process**

Solution design onboards the **claims handlers**, the **data scientists**, and the **IT team** for deployment at scale

Core work on developing algorithms requires **dedicated approaches to each specific task** for successful Proof of Concept

Case Study #2

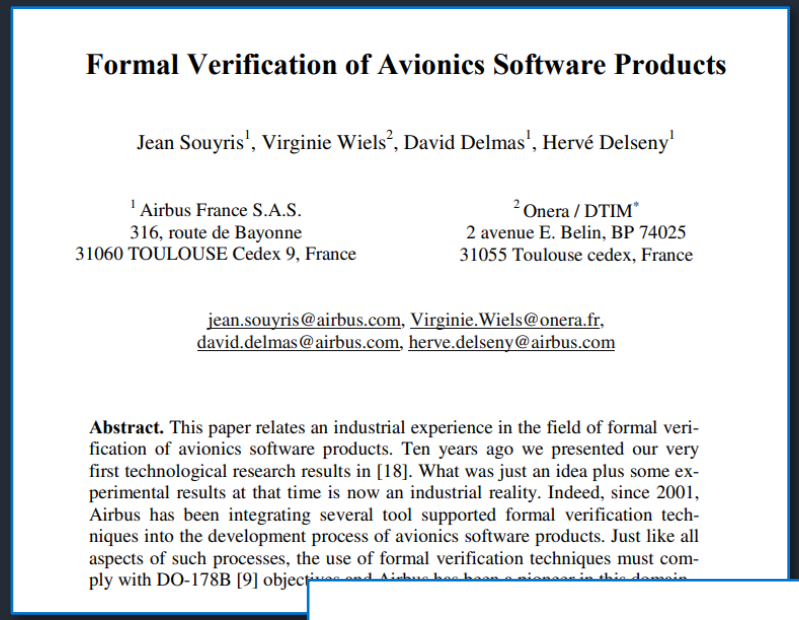
Building a Compliance System for Actuarial Models

Milliman

Best-in-class compliance standards: how aeronautics leads the way with the formal verification of critical systems

*"if a software product is developed by performing the activities prescribed by DO-178B successfully it will be considered as '**good for flight**' by the regulation authorities."*

Source : Souyris, J., Wiels, V., Delmas, D., & Delseny, H. (2009). Formal verification of avionics software products. In FM 2009: Formal Methods: Second World Congress, Eindhoven, The Netherlands, November 2-6, 2009. Proceedings 2 (pp. 532-546). Springer Berlin Heidelberg.



Formal Verification of Avionics Software Products

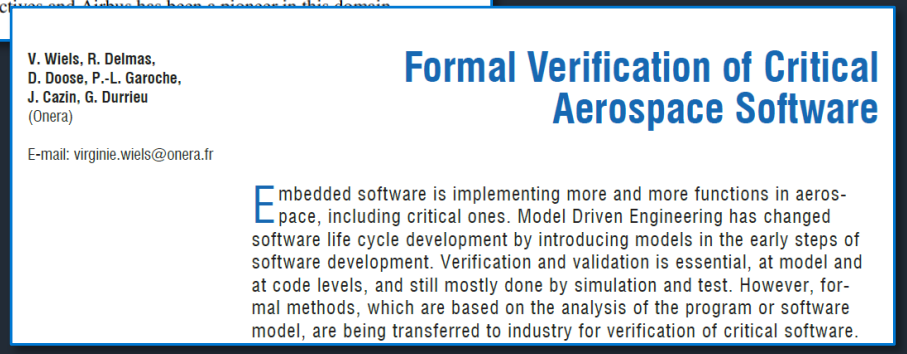
Jean Souyris[1], Virginie Wiels[2], David Delmas[1], Hervé Delseny[1]

[1] Airbus France S.A.S.
316, route de Bayonne
31060 TOULOUSE Cedex 9, France

[2] Onera / DTIM*
2 avenue E. Belin, BP 74025
31055 Toulouse cedex, France

jean.souyris@airbus.com, Virginie.Wiels@onera.fr,
david.delmas@airbus.com, herve.delseny@airbus.com

Abstract. This paper relates an industrial experience in the field of formal verification of avionics software products. Ten years ago we presented our very first technological research results in [18]. What was just an idea plus some experimental results at that time is now an industrial reality. Indeed, since 2001, Airbus has been integrating several tool supported formal verification techniques into the development process of avionics software products. Just like all aspects of such processes, the use of formal verification techniques must comply with DO-178B [9] objectives and Airbus has been a pioneer in this domain



V. Wiels, R. Delmas,
D. Doose, P.-L. Garoche,
J. Cazin, G. Durrieu
(Onera)

E-mail: virginie.wiels@onera.fr

Formal Verification of Critical Aerospace Software

Embedded software is implementing more and more functions in aerospace, including critical ones. Model Driven Engineering has changed software life cycle development by introducing models in the early steps of software development. Verification and validation is essential, at model and at code levels, and still mostly done by simulation and test. However, formal methods, which are based on the analysis of the program or software model, are being transferred to industry for verification of critical software.

Model verification in insurance relies on assessing its compliance against a set of prescribed specifications and regulations

An actuarial cash-flow model shall comply with a variety of unstructured information

Regulation

In the form of documents (ex: SII Delegated Acts)

Contracts

Describing how insurance products work and their parameters

Technical specifications

Specifying the rules and behaviour that the code shall reflect

Management rules

Coherence with the way the insurance company is managed in practice

While LLMs can structurally handle code, vision models are required to accurately comprehend actuarial specifications including formulas

LLM

Code

```
        Else ; cible epargne
            TARGET_CRED_RATE_PRD(Idx_Cflow) := MIN(MAX( TAUX_CONC , PrevCredRate + CIBLE_MINCRED_PC / 100 ), PrevCredRate + C
        EndIf

        ; bonus / malus de cible
        TARGET_CRED_RATE_PRD(Idx_Cflow) := TARGET_CRED_RATE_PRD(Idx_Cflow) + I_CFLOW_DEP.BONUS_TARGET_PC(Idx_Cflow)/100

        ;taux cible de l'assuré (pour calcul rachat dyn.)
        TARGET_CRED_RATE_POLICY_PRD(Idx_Cflow) := TARGET_CRED_RATE_PRD(Idx_Cflow)

        ; modification de la cible assureur en fonction des richesses latentes
        If I_CUSTOMCODE.BUFFER = 1 Then
            Idx_Pool := I_CFLOW_DEP.POOL(Idx_Cflow)
            If A_POOL_RULES.LG(Idx_Pool) = 0 And L_DET_CFLOWS(t).MATH_RES_BR_PL(Idx_Pool)>0 Then
                Ressources_PL := max( 0 , (L_ADD_RES_ST(t).FUND_PB_PL_ST(Idx_Pool) + F_EQUIT_BD(t).EQUIT_MV_PL(Idx_Pool) - F_E

                ; sous Seuil_MIN_1_PC, target min / au dela de Seuil_MAX_1_PC, target max
                ; entre Seuil_MIN_1_PC et Seuil_MIN_2_PC, convergence vers target
                ; entre Seuil_MIN_2_PC et Seuil_MAX_2_PC, pas d'impact (=target)
                ; entre Seuil_MAX_2_PC et Seuil_MAX_1_PC convergence vers target max

                ; test d'erreur sur le paramétrage du buffer
                If Idx_Cflow = 1 And t = SYS_CALCSTEP Then
                    If (Buffer_Factor_Min > 1) Or (Buffer_Factor_Max < 1) Or (Buffer_Factor_Max > 100) Or (Buffer_Factor_Max <
l_MAX_1_PC) Or (Seuil_MIN_2_PC > Seuil_MAX_2_PC) Or (Seuil_MAX_1_PC<1) Or (Seuil_MAX_2_PC<1) Then
                        warning("anomalie dans le parametrage du buffer")
                    EndIf
                EndIf

                Target_MIN := Buffer_Factor_Min * TARGET_CRED_RATE_PRD(Idx_Cflow)
                Target_MAX := Buffer_Factor_Max * TARGET_CRED_RATE_PRD(Idx_Cflow)
                Target     := 1                 * TARGET_CRED_RATE_PRD(Idx_Cflow)

                If Ressources_PL <= Seuil_MIN_1_PC/100 Then
                    TARGET_CRED_RATE_PRD(Idx_Cflow) := Target_MIN
```

Documentation

Moyenne des Taux Longs

La moyenne des taux longs sur un nombre d'années défini par `CIBLE_ML_YR` est calculée comme la somme des taux pondérés par leur poids respectif, divisée par le nombre d'années:

$$\text{Moy\_Taux\_Longs} = \frac{\text{Taux\_Long\_N} + \text{Taux\_Long\_N1} + \ldots + \text{Taux\_Long\_N7}}{\text{CIBLE\_ML\_YR}}$$

Taux Concurrentiel

Le taux concurrentiel (`TAUX_CONC`) est calculé comme le maximum entre trois composantes: le produit du taux court et d'un coefficient, le produit du taux long et d'un coefficient, et le produit de la moyenne des taux longs pondérée avec la performance des actions, chaque composante étant ajustée par un coefficient spécifique.

$$\text{TAUX\_CONC} = \max\left(\frac{\text{CIBLE\_TC\_PC}}{100} \times \text{Taux\_Court\_N}, \frac{\text{CIBLE\_TL\_PC}}{100} \times \text{Taux\_Long\_N}, \frac{\text{CIBLE\_ML\_PC}}{100} \times \left(\frac{\text{POND\_RATES\_TL\_PC}}{100} \times \text{Moy\_Taux\_Longs} + \frac{\text{POND\_IDX1\_PC}}{100} \times \text{Perf\_1}\right)\right)$$

Ensuite, des ajustements sont appliqués pour obtenir le taux concurrentiel net de chargements.

Top-tier vision models can comprehend actuarial formulas, which can then **feed the LLM for code verification**

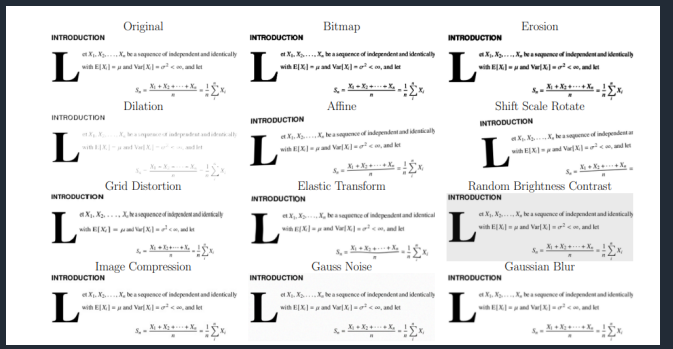
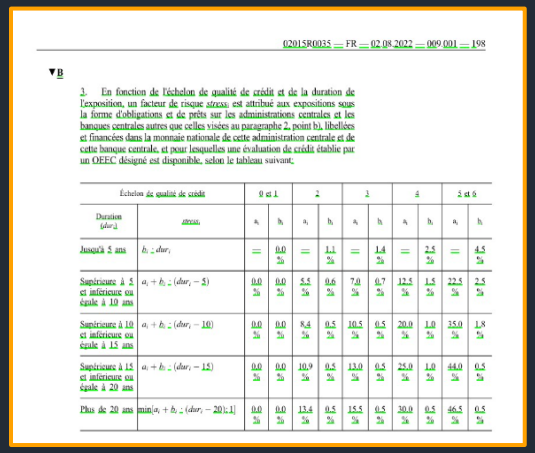


Illustration: how to check consistency with external unstructured information – example SCR Spread from Delegated Acts

Delegated Acts are provided as pdf



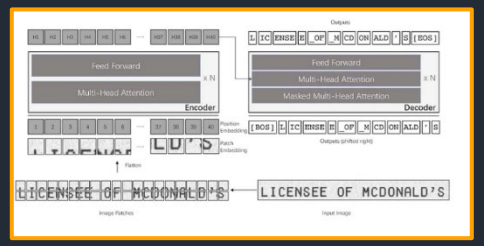
02015R0035 — FR — 02.08.2022 — 009.001 — 198

▼B

3. En fonction de l'échelon de qualité de crédit et de la duration de l'exposition, un facteur de risque «stress» est attribué aux expositions sous la forme d'obligations et de prêts sur les administrations centrales et les banques centrales autres que celles visées au paragraphe 2, point b), libellées et financées dans la monnaie nationale de cette administration centrale et de cette banque centrale, et pour lesquelles une évaluation de crédit établie par un OEEC désigné est disponible, selon le tableau suivant:

Échelon de qualité de crédit		0 et 1		2		3		4		5 et 6	
Duration (dur_i)	stress:	a_i	b_i	a_i	b_i	a_i	b_i	a_i	b_i	a_i	b_i
Jusqu'à 5 ans	$b_i \cdot dur_i$	=	0,0 %	=	1,1 %	=	1,4 %	=	2,5 %	=	4,5 %
Supérieure à 5 et inférieure ou égale à 10 ans	$a_i + b_i \cdot (dur_i - 5)$	0,0 %	0,0 %	5,5 %	0,6 %	7,0 %	0,7 %	12,5 %	1,5 %	22,5 %	2,5 %
Supérieure à 10 et inférieure ou égale à 15 ans	$a_i + b_i \cdot (dur_i - 10)$	0,0 %	0,0 %	8,4 %	0,5 %	10,5 %	0,5 %	20,0 %	1,0 %	35,0 %	1,8 %
Supérieure à 15 et inférieure ou égale à 20 ans	$a_i + b_i \cdot (dur_i - 15)$	0,0 %	0,0 %	10,9 %	0,5 %	13,0 %	0,5 %	25,0 %	1,0 %	44,0 %	0,5 %
Plus de 20 ans	$\min[a_i + b_i \cdot (dur_i - 20); 1]$	0,0 %	0,0 %	13,4 %	0,5 %	15,5 %	0,5 %	30,0 %	0,5 %	46,5 %	0,5 %

Vision model allows the extraction of all the source information (descriptions, tables, coefficients, formulas, etc.)



The Large Language Model recovers an understanding of how the spread shock calculation works

Le tableau indique pour la qualité de crédit 3 et la tranche de duration supérieure à 10 ans et inférieure ou égale à 15 ans, les valeurs de (ai) et (bi) sont respectivement de 10,5% et 0,5%.

Par conséquent, le facteur de stress (stress_i) pour une obligation de qualité de crédit 3 et de duration 12 ans est calculé comme suit:

$$stress_i = ai + bi \cdot (duri\text{–}10) = 10,5\% + 0,5\% \cdot (12\text{–}10) = 10,5\% + 0,5\% \cdot 2 = 10,5\% + 1\% = 11,5\%.$$

Le facteur de stress à appliquer est donc de 11,5%.

The system is operational to produce a mirror calculation of shocked market values and to perform checks on the calculation chain

Spread SCR calculation per model point

1	GOV_UE	AAA	11.0	613 802.7	-	0.00%	-
59	GOV_UE	AAA	14.7	8 841.3	-	0.00%	-
60	CORPORATE	AAA	6.2	614 603.0	31 432.90	5.11%	31 432.90
61	CORPORATE	AAA	9.1	204 962.8	13 377.70	6.53%	13 377.70
67	CORPORATE	AAA	8.1	10 854.0	655.89	6.04%	655.89
89	CORPORATE	BBB	6.0	2 325 964.1	324 699.31	13.96%	324 699.31
90	CORPORATE	BBB	8.7	605 977.0	109 069.53	18.00%	109 069.53

These model verification approaches can be extended to other areas that include unstructured information: technical specifications, internal guidelines, contractual clauses, distribution agreements, reinsurance treaties, etc.

A client project focused on Excel files processing required to develop a new way to deliver spreadsheet information to the LLM

Why processing spreadsheets with LLMs?

This offers advanced capabilities in understanding, manipulation, and question/answering

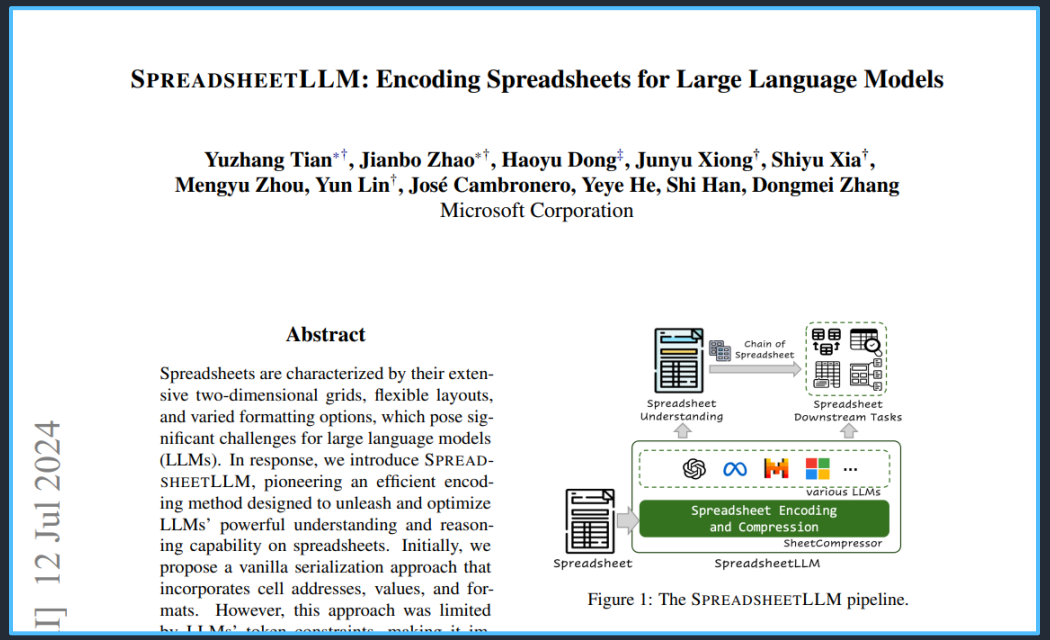
Challenges?

Complexity of managing large two-dimensional grids, complex layouts, cell addresses and formats… with, on top, token limits typically exceeded

The trick

An innovative data encoding mechanism that allows us to compress information while preserving its structure

Concepts related to spreadsheet encoding for LLM are inspired from the paper below:



SPREADSHEETLLM: Encoding Spreadsheets for Large Language Models

Yuzhang Tian*†, Jianbo Zhao*†, Haoyu Dong‡, Junyu Xiong†, Shiyu Xia†,
Mengyu Zhou, Yun Lin†, José Cambronero, Yeye He, Shi Han, Dongmei Zhang
Microsoft Corporation

Abstract

Spreadsheets are characterized by their extensive two-dimensional grids, flexible layouts, and varied formatting options, which pose significant challenges for large language models (LLMs). In response, we introduce SPREADSHEETLLM, pioneering an efficient encoding method designed to unleash and optimize LLMs' powerful understanding and reasoning capability on spreadsheets. Initially, we propose a vanilla serialization approach that incorporates cell addresses, values, and formats. However, this approach was limited

Figure 1: The SPREADSHEETLLM pipeline.

Case Study #3

Automating the Production of Reports

Milliman

Analysis of the Solvency II Own Funds movement between Q4 23 and Q1 24

Focus on the Own Funds for the Euro contracts

Summary

The **Own Funds for Euro contracts closed at €1,858.1 million**, representing a **significant decrease of -€333.0 million** (-15.2%). This variation is mainly explained by the following elements:

- The impact of **updating RC and PPE provisions for -€252.1 million** leading to:
  - A decrease in the terminal value of RC recognized for the shareholder by -€228.4 million, and
  - A margin effect of -€16.9 million, mainly explained by the increase in own fund costs.
- The impact of **-€146.8 million due to the increase in spreads** (AAA 10 years +12.10 bps) leading to:
  - A margin effect explained by the increase in own funds costs of -€116.6 million (approximately -0.02cts/year), and
  - A duration effect explained by the increase in own funds costs of -€11.2 million.

Level 1 analysis: which main driver contributes to the variation?

- The impact of **-€101.6 million due to the increase in the swap rate** (10 years +7.7 bps) leading to:
  - A margin effect explained by the increase in own funds costs of -€79.5 million (approximately -0.01cts/year), and
  - A duration effect explained by the increase in own funds costs of -€22.1 million.
- The **impact of equity markets for +€180.5 million** (Eurostoxx performance of +12.4%) leading to:
  - A margin effect of +€143.5 million, mainly explained by the increase in loading capacity on financial production of +€67 million and the decrease in own funds costs of +€58.5 million, and
  - A duration effect of +€23.0 million following the increase in duration by +0.2% from 9.3 to 9.5 years.

Level 2 analysis: why?

- A duration effect of -€10.6 million following the decrease in duration by -0.5% from 9.3 to 9.2 years.

Analysis of movement

Step	OF (m€)	Variation
Opening	2 191.1	
Model change	2 191.1	0.0
New Business	2 148.6	-2.0
Lapse update	2 097.3	-51.3
Equity	2 277.7	180.5
Interest rate	2 176.1	-101.6
Interest rate volatility	2 221.2	45.1
Inflation	2 208.0	-13.3
Diversified	2 257.0	49.0
Spreads	2 110.2	-146.8
Update RC & PPE	1 858.1	-252.1
Closing	1 858.1	-333.0

The expert refers to economic data to rationalise the observation made

The expert typically computes indicators « in parallel » to understand the contribution of each effect

Why expertise is essential to “constrain” LLM to follow expert reasoning

“[…] current LLMs are not capable of genuine logical reasoning; instead, they attempt to replicate the reasoning steps observed in their training data.”

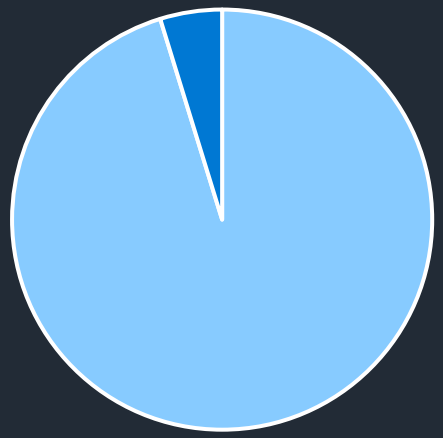
— Mirzadeh, I., Alizadeh, K., Shahrokhi, H., Tuzel, O., Bengio, S., & Farajtabar, M. (2024). Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models.

Automatically generated reporting for the validation of economic scenarios

LLMs revolutionise **large-scale analysis** by extracting key insights and following (constrained) expert reasoning to **craft flexible, synthesised narratives**

Return on Investment through improved efficiency, higher quality, and responsiveness

95% reduction in reporting production time



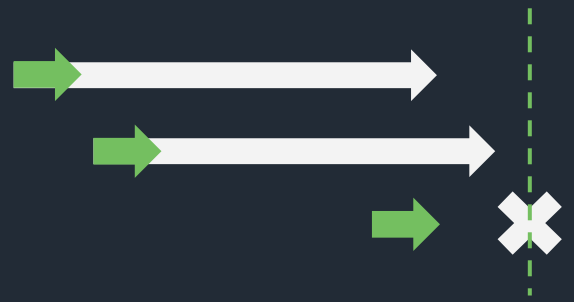
Higher employee satisfaction and workforce for high value projects

Systematic expert analyses and reasoning

"A lower repricing accuracy has been noted this quarter, mainly due to outliers detected in swaption market data. Overall, repricing tests remain still valid since the average repricing error remains in the range of scenario productions from last year."

Increased quality of reports and reduced "key person risk"

Ability to deliver reporting for short notice economic scenarios requests



Augmented visibility and support from management

Case Study #4

Delivering Chatbots for Efficient Knowledge Accessibility

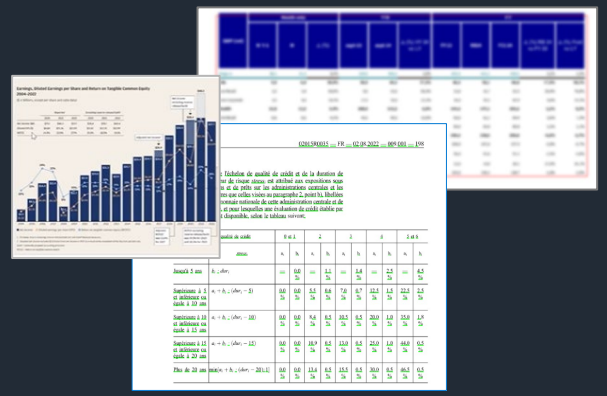
Milliman

Developing chatbots requires solving key technical challenges, such as information retrieval, visualisation, computation, and reasoning

Intent detection layers combined with agentic workflows allow for an efficient user / chatbot interaction

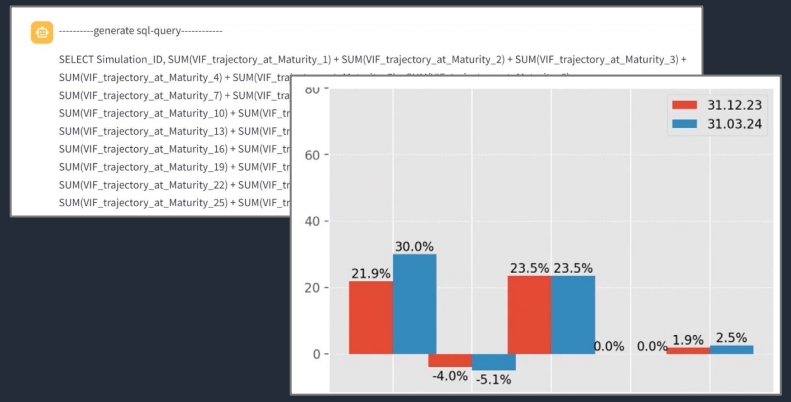
Information retrieval

Development and/or integration of a collection of vision models to **retrieve and understand information in a variety of modalities (tables, graphs, formulas)**



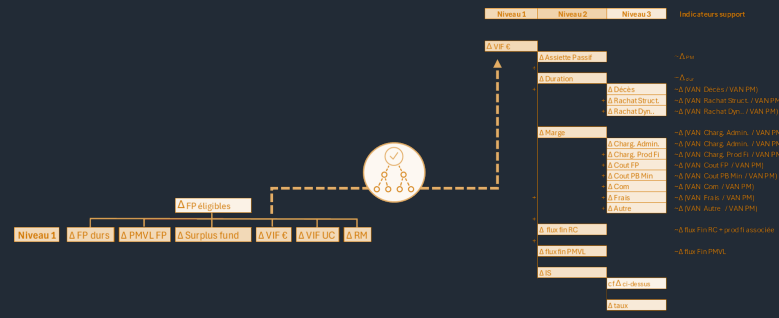
Computation

Development of “agent” capabilities for LLM to generate the code (ex: Python) needed to **perform the calculations** required to respond to a user request



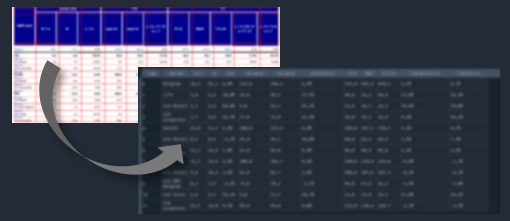
Actuarial reasoning

Incorporation of the actuarial reasoning into LLM. It takes the form of an analysis by layers that replicates, e.g., the **actuarial analysis of the movement of the solvency ratio over a period of interest**

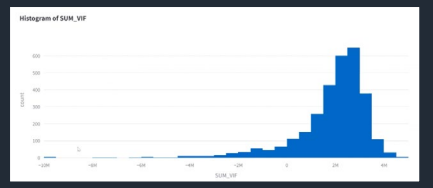


Delivering instantaneous knowledge has driven value in the areas of actuarial work, expertise, software, and regulation

A project aimed to overcome performance issues on **information retrieval within typical actuarial documents** for a chatbot intended for actuaries



Chatbot to **access inputs/results of ALM cash-flow model** and perform analysis on the **evolution of the solvency ratio**



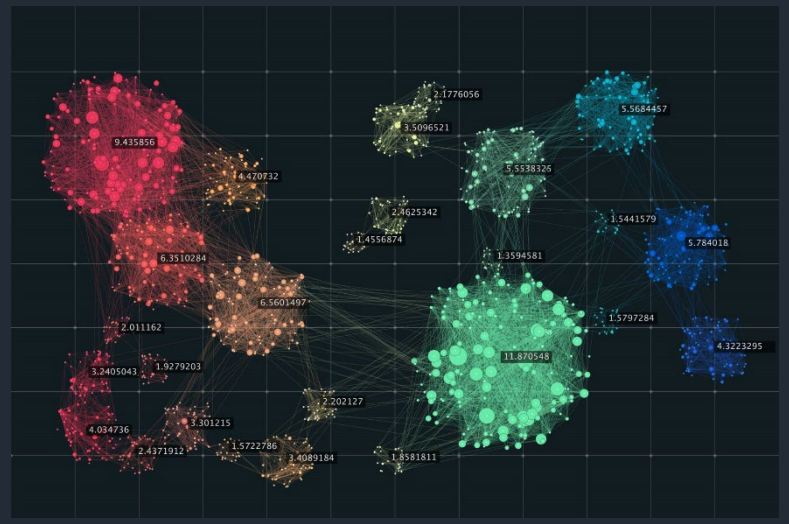
Supporting users of **actuarial software** to navigate within user guides and model specifications

Chatbot on the **European regulation**, that preserves the integrity and understanding of the information (tables, formulas, values)

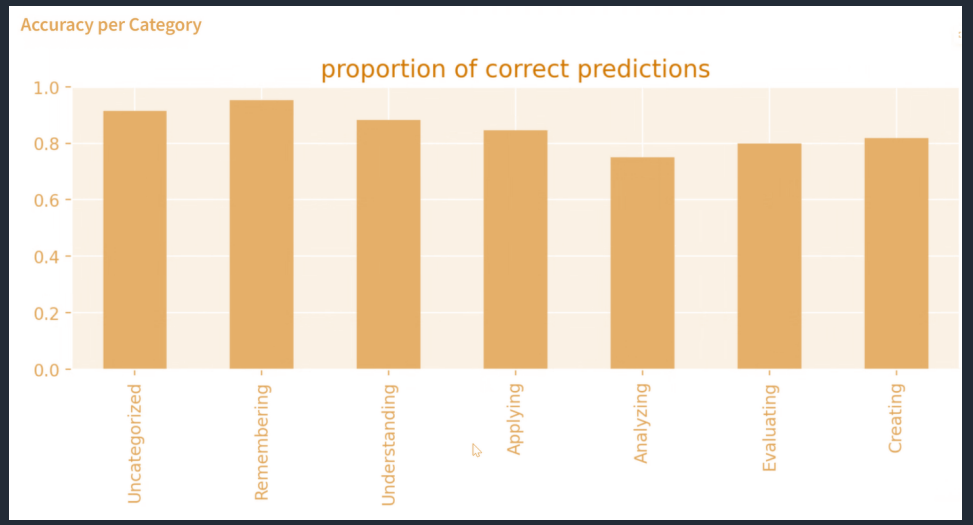
Échelon de qualité de crédit		0 et 1		2		3		4		5 et 6	
Duration (dur_i)	stress$_i$	a_i	b_i	a_i	b_i	a_i	b_i	a_i	b_i	a_i	b_i
Jusqu'à 5 ans	$b_i \cdot dur_i$	=	0.0 %	=	1.1 %	=	1.4 %	=	2.5 %	=	4.5 %
Supérieure à 5 et inférieure ou égale à 10 ans	$a_i + b_i \cdot (dur_i - 5)$	0.0 %	0.0 %	5.5 %	0.6 %	7.0 %	0.7 %	12.5 %	1.5 %	22.5 %	2.5 %

Measuring and monitoring performance can be achieved by LLM evaluation pipelines, combining topic modelling with accuracy measurement

User queries classified per topic



Performance measure per topic



Case Study #5

Turning Unstructured Data into Insights

Milliman

An example of a geopolitical risk index to demonstrate how LLMs can convert text into quantitative insight, based on expert guidance


The New York Times.

U.S. AND SOVIET REACH ACCORD ON CUBA; KENNEDY ACCEPTS KHRUSHCHEV PLEDGE TO REMOVE MISSILES UNDER U.N. WATCH

DE GAULLE WINS 66% OF BALLOTS; ABSTENTIONS HIGH

TYRANT SEES VISIT; RUSSIAN ACCEPTS

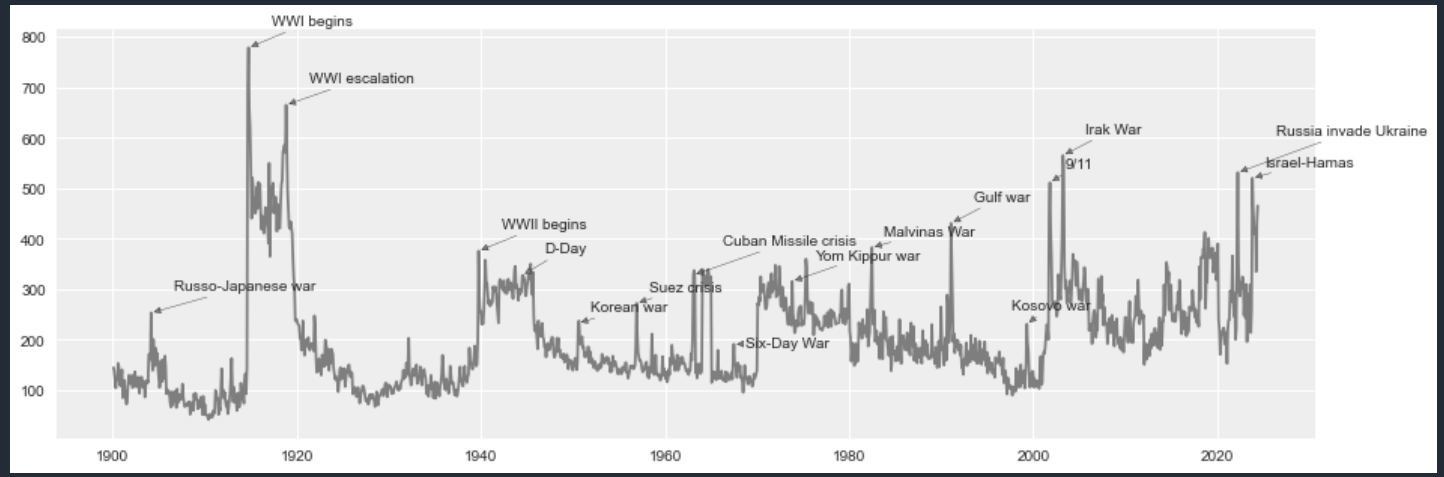
CAPITAL HOPEFUL

Castro Asks Guantanamo, Strives to Repair Prestige

Plane Veers Over Soviet, Kennedy Voices Regrets

New York Times: Historical news articles and headlines from the New York Times API from 1900 to 2024, that forms a **10+ million articles dataset**

After fine-tuning a model to capture geopolitical dynamics (+1 tension, -1 decrease), we built a **geopolitical risk sentiment index** that can then be fed by upcoming news in real time



Examples of use cases leveraging sentiment analysis in insurance and risk assessment

Customer satisfaction

Vox.IA (Covéa) analyses customer feedback from satisfaction surveys. It is operational since March 2023 and processes nearly 500,000 reviews annually.(1)

Mortality prediction

Sentiment scores measured in nursing notes are associated with the risk of 30-day mortality in sepsis patients and may improve the preview of 30-day mortality.(2)

Cyber risk monitoring

EIOPA delivers a text analysis based cyber risk indicator, calculated from earning calls transcripts from listed insurers.(3)

(1) https://www.covea.com/fr/analyse/lintelligence-artificielle-service-assures
(2) https://onlinelibrary.wiley.com/doi/pdf/10.1155/2021/1713363
(3) https://www.eiopa.europa.eu/tools-and-data/insurance-risk-dashboard_en

The GenAI Checklist

Milliman

The “Gen AI” checklist: spot high value areas for Gen AI

- ☐ **G**enius-level automation
- ☐ **E**quivalence for compliance
- ☐ **N**arrative reports
- ☐ **A**ccessible knowledge
- ☐ **I**nsightful data

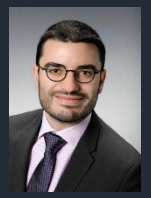
☒

If you wish to check at least one box, then (Gen) AI is the right choice for your business!

Q&A

Paul Marron

Paul.Marron@milliman.com



Alexandre Boumezoued

Alexandre.Boumezoued@milliman.com

Milliman

Solutions for a world at risk™

Milliman

Thank you

Paul Marron

Paul.Marron@milliman.com

Alexandre Boumezoued

Alexandre.Boumezoued@milliman.com