Advanced Applications of Generative AI in Actuarial Science: Case Studies Beyond ChatGPT

Iris Nonneman (Achmea) & Simon Hatzesberger (Deloitte)

6th European
Congress of Actuaries
www.eca2026.org

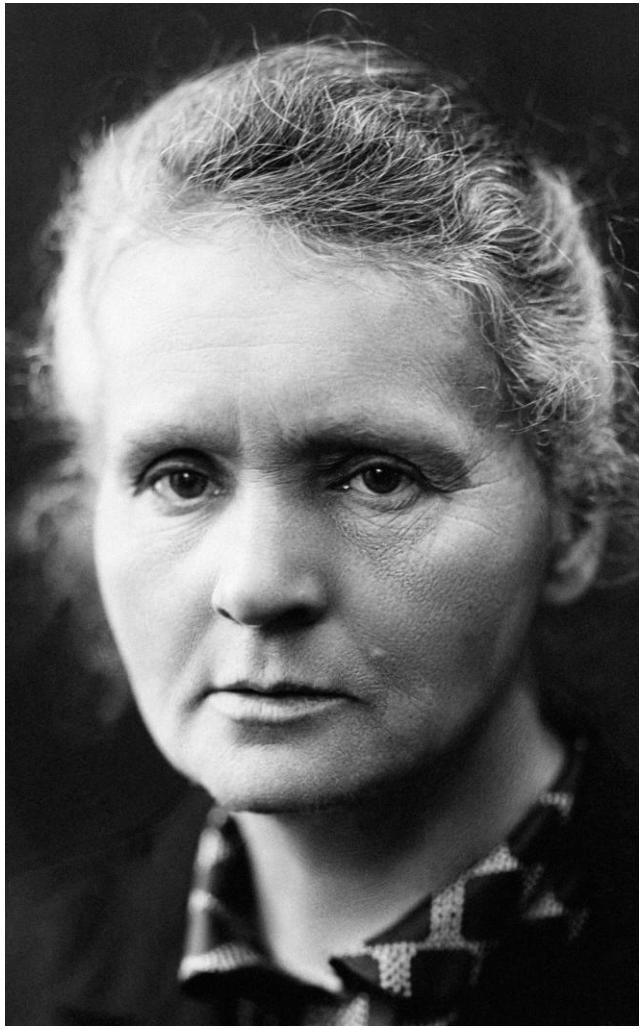
Agenda

- History and Introduction of Advanced AI
- Four Practical GenAI Case Studies
  - → Case Study 1: Improving Claim Cost Prediction with LLM-Features
  - → Case Study 2: GenAI-Driven Market Comparison
  - → Case Study 3: Car Damage Classification with Vision-Enabled LLMs
  - → Case Study 4: Multi-Agent System for Legacy Code Migration
- Further GenAI Applications in the Insurance Industry
- Q&A

Discovering the possibilities of AI

‘Nothing in life is to be feared, it is only to be understood. Now is the time to understand more, so that we may fear less.’

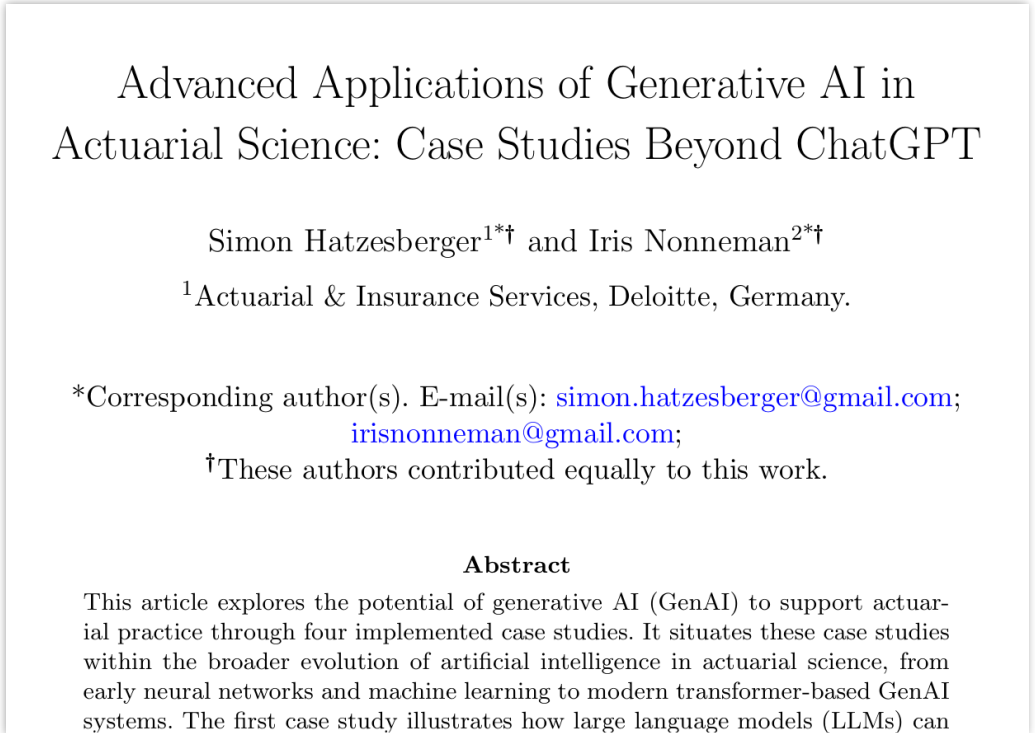
Marie Curie



https://www.nationalgeographic.com/culture/article/131009-why-we-care-about-nobel-history-curie (photograph from Hulton-Deutsch collections/corbis

Supplementary Material to This Presentation

More content available in the article “Advanced Applications of Generative AI in Actuarial Science: Case Studies Beyond ChatGPT” and all implementations on the GitHub Actuarial AI Case Studies repository of the IAA AI Task Force. The article also discusses the ethical and technical challenges of deploying GenAI in actuarial work.



Advanced Applications of Generative AI in Actuarial Science: Case Studies Beyond ChatGPT

Simon Hatzesberger[1*†] and Iris Nonneman[2*†]

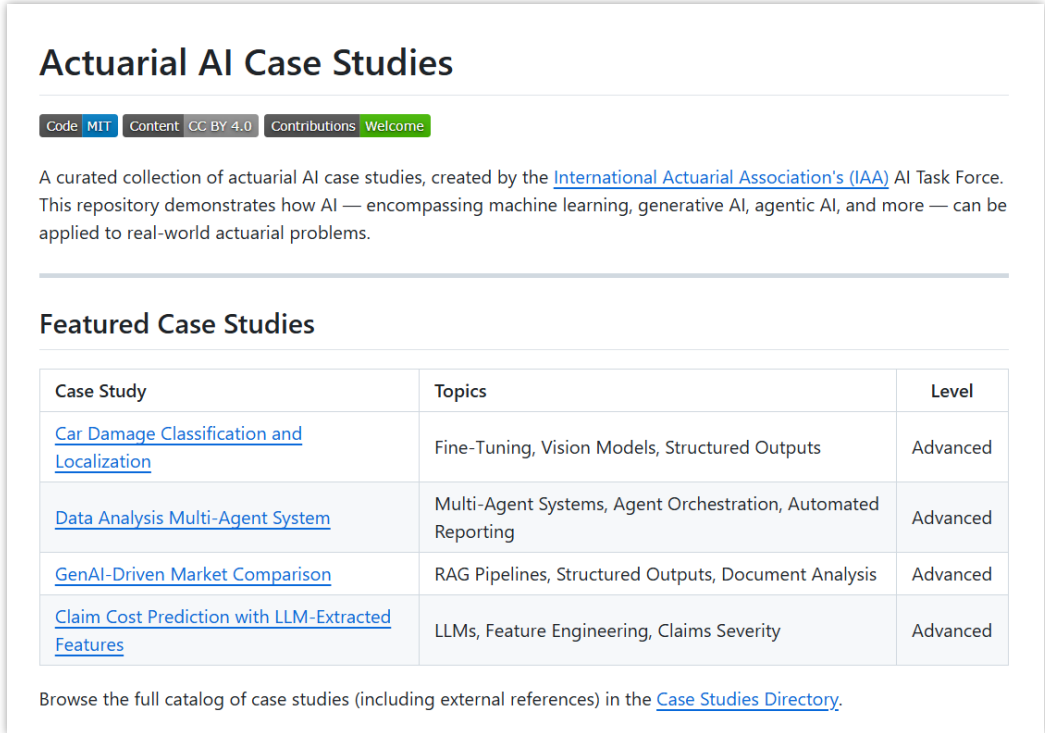
[1]Actuarial & Insurance Services, Deloitte, Germany.

*Corresponding author(s). E-mail(s): simon.hatzesberger@gmail.com; irisnonneman@gmail.com;
†These authors contributed equally to this work.

Abstract

This article explores the potential of generative AI (GenAI) to support actuarial practice through four implemented case studies. It situates these case studies within the broader evolution of artificial intelligence in actuarial science, from early neural networks and machine learning to modern transformer-based GenAI systems. The first case study illustrates how large language models (LLMs) can

https://arxiv.org/abs/2506.18942



Actuarial AI Case Studies

Code MIT Content CC BY 4.0 Contributions Welcome

A curated collection of actuarial AI case studies, created by the International Actuarial Association's (IAA) AI Task Force. This repository demonstrates how AI — encompassing machine learning, generative AI, agentic AI, and more — can be applied to real-world actuarial problems.

Featured Case Studies

Case Study	Topics	Level
Car Damage Classification and Localization	Fine-Tuning, Vision Models, Structured Outputs	Advanced
Data Analysis Multi-Agent System	Multi-Agent Systems, Agent Orchestration, Automated Reporting	Advanced
GenAI-Driven Market Comparison	RAG Pipelines, Structured Outputs, Document Analysis	Advanced
Claim Cost Prediction with LLM-Extracted Features	LLMs, Feature Engineering, Claims Severity	Advanced

Browse the full catalog of case studies (including external references) in the Case Studies Directory.

https://github.com/IAA-AITF/Actuarial-AI-Case-Studies/

History and Introduction of Advanced AI

AI Taxonomy

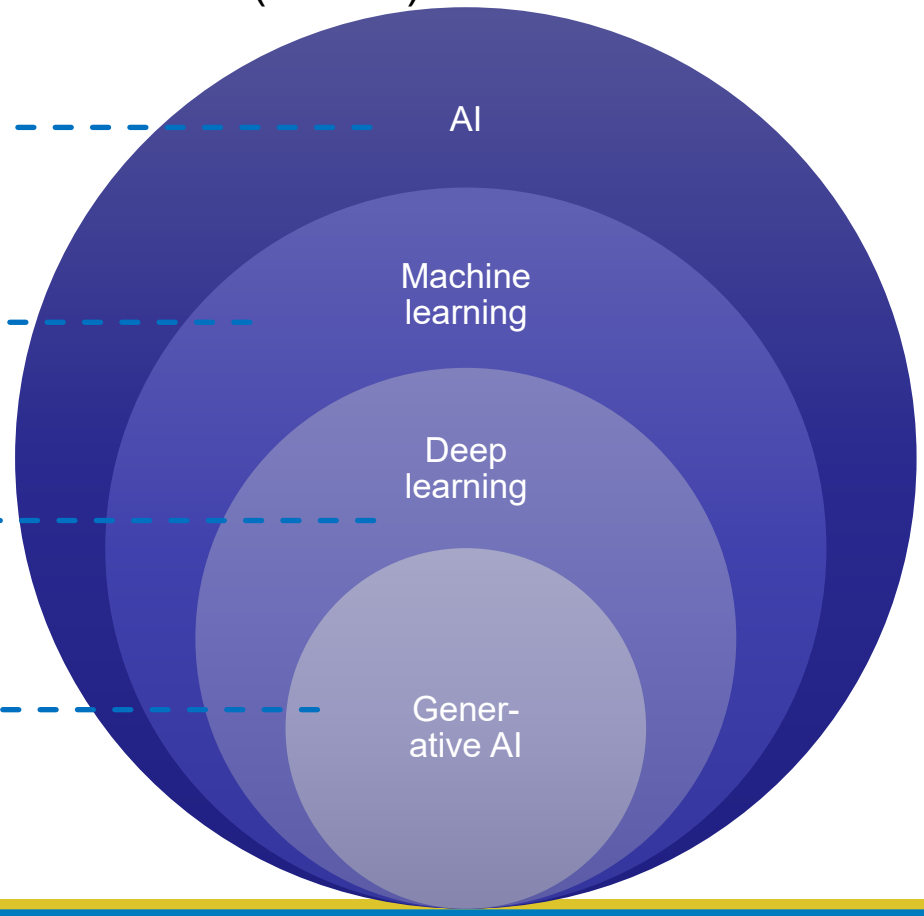
Artificial intelligence system (AI system) is a machine-based system that is designed to operate with varying levels of *autonomy* and that can, for explicit or implicit objectives, *generate output* such as predictions, recommendations, or decisions influencing physical or virtual environments (OECD).

AI: development of smart systems that can carry out tasks typically requiring human intelligence

Machine learning: subfield of AI, uses algorithms that can learn from data and make accurate predictions

Deep learning: subfield of ML, applies neural nets to learn complex patterns

Generative AI: applies transformer neural nets to generate data



History of AI *and Actuaries*

1960+

AI originated in the late 20th century with advances in computational intelligence.

Constrained by computational resources

Actuaries: mainly rely on analytical methods, e.g.:

- GLM
- Bornhuetter-Ferguson

1980

breakthroughs with backpropagation and neural nets.

2010

Actuaries: start to experiment with ML techniques

marked rapid progress with deep learning, AlexNet, and transformer models.

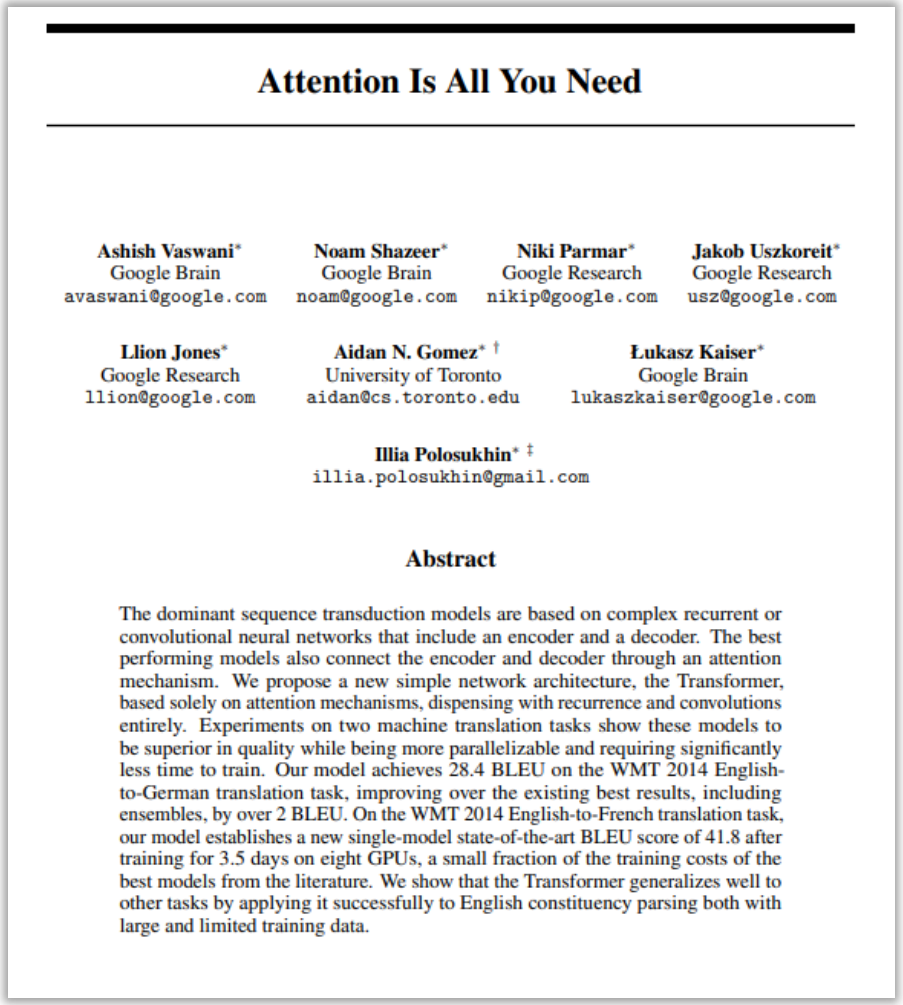
Advances of computing power and data availability

2017

Transformer neural nets open the gates for GenAI, automation of processes starts to take off

Adoption of NNs and GenAI

The Rise of Pattern Recognition



Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

- *Attention Is All You Need* introduced transformer networks
- Emergence of LLMs
- GenAI is on the rise in business since the release of ChatGPT (Nov 2022)

Four Practical GenAI Case Studies

Improving Claim Cost Prediction with LLM-Features

Setting & Problem

Traditional models ignore valuable **unstructured text data** in claim descriptions

Solution

Dataset: 3,000 workers' compensation claims with both tabular and textual data

LLM Approach:

- Extract structured features from claim descriptions using targeted prompts
- Focus on number of body parts injured, main body part, cause of injury

Model Enhancement: Compare baseline gradient boosting vs. enhanced model with LLM features

The LLM Prompt – Structured Output Extraction

SYSTEM PROMPT

```
Your task is to extract structured
information about injuries.

number_of_body_parts_injured:
  Total count of injured parts.

main_body_part_injured:
  Primary area (e.g. "HEAD", "THUMB").

cause_of_injury:
  Primary action verb ("fall", not
  "fell from box").
```

LLM OUTPUT EXAMPLE

Input claim:

"Claimant slipped on wet floor, injured lower back while falling."

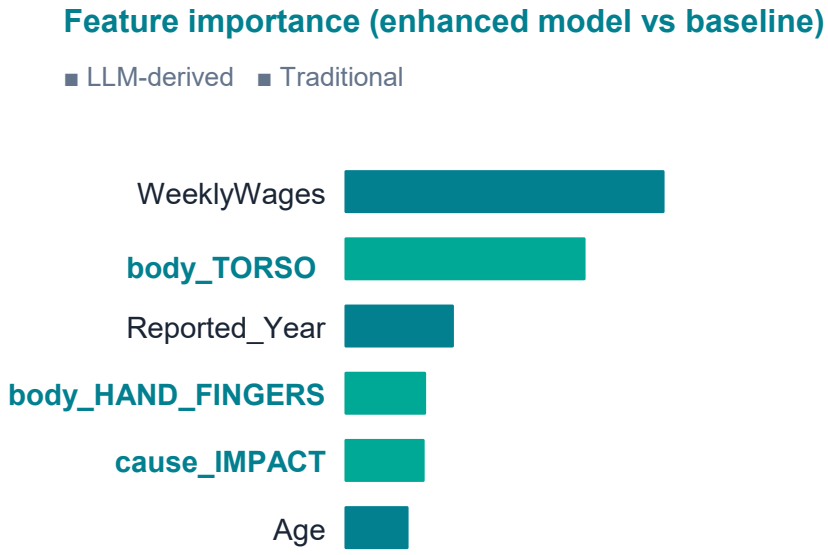
Structured output:

```
n_body_parts:     1
main_body_part:   "LOWER BACK"
cause_of_injury:  "slip"
```

Approach

1
Baseline Model
Structured tabular features only (age, wages, working days)
2
LLM Extraction
LLM extracts body part, cause, count
3
Categorisation
8 body regions + 13 cause categories
4
Enhanced Model
Baseline + LLM features Gradient Boosting Regressor
5
Evaluation
CV + held-out test RMSE, MAE, R², quantile loss

Feature Importance & Key Takeaways



Key Takeaways

- LLM features significantly boost all metrics across all claim sizes
- Body part > cause > count in terms of individual contribution
- Effect is stable over 50 cross-validation folds (low SD)
- LLM output may vary across runs → cache raw outputs

GenAI-Driven Market Comparison (1/2)

Setting & Problem

Market comparisons involve extracting and analysing key data – such as from annual reports or insurance tariffs – to uncover strategic insights across competitors.

→ **Manual Comparisons Fall Short**
Traditional methods are slow, error-prone, and not scalable for diverse, unstructured data.

→ **Strategic Blind Spots**
Inefficient processes delay decision-making and limit competitive insight.

Solution

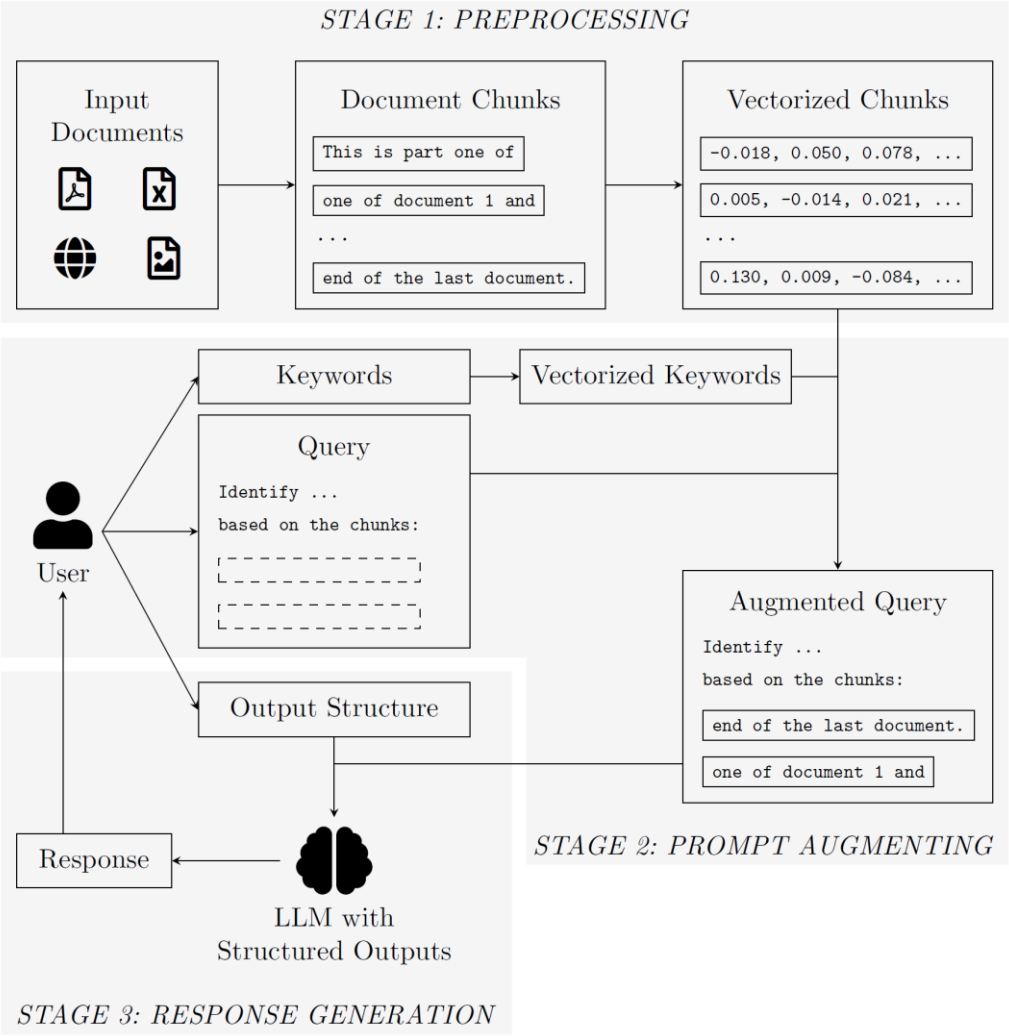
→ **GenAI-Powered Automation**
Retrieval-Augmented Generation (RAG) and Structured Outputs enable efficient extraction and comparison of key data at scale.

→ **Cross-Domain Applicability**
Solution applies to annual reports, tariff comparisons, etc.

GenAI-Driven Market Comparison (2/2)

Key Results:

- ➜ Benchmarked five LLMs on three extraction aspects across AXA, Generali and Zurich (900 extractions in total).
- ➜ Validated against a manually-extracted ground truth of 45 fields from the 2025 annual reports.
- ➜ GPT-4.1 and GPT-5.4 passed 100% of runs on every aspect; solvency ratios and discount rates were reliable across all five models.
- ➜ Financial strength ratings were the hardest aspect (66.7% to 100%): entity-level ambiguity and IFSR vs. issuer/debt ratings need careful prompting.



Car Damage Classification with Vision-Enabled LLMs (1/2)

Setting & Problem

Motor insurers receive **large volumes of photo-based car damage claims** that need **fast, consistent triage and classification**.

→ **Traditional CNN models** achieve good accuracy on damage type but **capture no to little contextual information** (e.g. damage location, severity, lighting, weather).

→ Actuaries and claims handlers lack structured, machine-readable image outputs to support pricing, claims routing, and fraud detection.

Solution

→ **Vision-Enabled LLMs, Fine-Tuning & Structured Outputs**
Use GPT-4o fine-tuned on a dataset of thousands of labelled car damage images, classified into six predefined damage types.

→ **Extraction of Contextual Information**
Extend the prompt so the model predicts the damage class and outputs structured context (e.g. damage location) for use in claims and actuarial workflows.

Car Damage Classification with Vision-Enabled LLMs (2/2)

Key Result 1:

- ➜ Non-fine-tuned vision-enabled LLMs already perform well out of the box
- ➜ Fine-tuning boosts classification performance

Prediction Model	Accuracy (↑)	Weighted F1 Score (↑)
Convolutional Neural Network	0.837	0.835
Non-Fine-Tuned GPT-4o	0.823	0.825
Fine-Tuned GPT-4o	0.880	0.880

Key Result 2:

- ➜ Beyond pure damage classification, vision-enabled LLMs can also capture rich visual context (e.g. damage location).

Actual damage type	glass shatter	tire flat	dent
Predicted damage type	glass shatter	tire flat	dent
Predicted damage location	windshield	—	rear bumper

Multi-Agent System for Legacy Code Migration (1/2)

Setting & Problem

Critical actuarial logic is often locked in legacy languages (R, SAS, COBOL, VBA) that are costly to maintain and risky to migrate by hand.

→ Migrate this code to modern Python without losing the audit trail.

→ Validate every migration against a prespecified test suite.

Solution

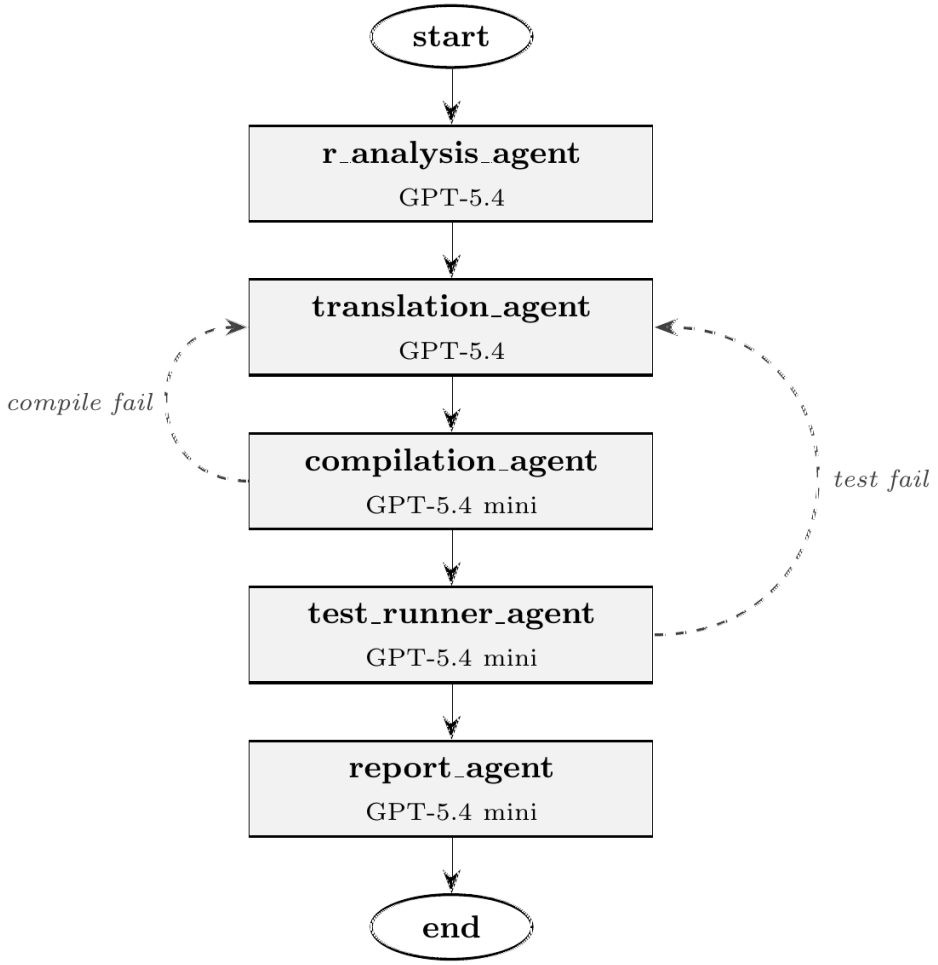
→ **Five specialised LLM agents orchestrated by LangGraph:**
analyse → translate → compile → test → report
GPT-5.4 for the heavy lifting (analysis + translation), GPT-5.4 mini for compilation, testing & reporting.

→ **Test failures loop back to the translator for targeted fixes**
R-derived ground truth is held outside the translator's context — the system cannot succeed by injecting the answer.

Multi-Agent System for Legacy Code Migration (2/2)

Key Results:

- ➜ Validated on two actuarial reserving tasks: chain-ladder (198 lines of R) and GLM with bootstrap (265 lines of R).
- ➜ 100% test pass rate across 10 end-to-end runs per task, against pre-written, R-verified suites of 14 and 15 tests.
- ➜ Self-correcting design: failed compilations or tests loop back to the translator (on average 1.2 retries for the harder GLM task, up to 4).
- ➜ Deterministic outputs reproduced to floating-point precision; every run produced functionally equivalent Python.



Further GenAI Applications in the Insurance Industry

Further Applications of Generative AI

Automated Reporting

Use prompt engineering, fine-tuned models, and tools to automate accurate, consistent regulatory reports (e.g., SFCRs).

Customer Interaction and Support

LLM chatbots with audio/vision understand text, voice, handwriting; deliver personalised policy help.

Claims Processing

Digitise, classify, and automate claims documents to speed settlements and cut end-to-end operating costs.

Underwriting Support

Analyse applicant data from multiple sources and generate structured risk summaries to speed up and standardise risk evaluation.

Employee Training

Create tailored learning materials and LLM chatbots to train actuaries and share knowledge.

...

Q&A

Speakers

Iris Nonneman

achmea

irisnonneman@gmail.com

Iris Nonneman is a data scientist, certified actuary (AAG), and yoga teacher. She works as a senior data scientist at Achmea. She participates in various working groups on actuarial data science and leads the subworkstream on developing AI case studies within the AI Task Force of the International Actuarial Association. She regularly speaks at conferences and provides workshops on AI.

Dr. Simon Hatzesberger

Deloitte.

shatzesberger@deloitte.de

Simon Hatzesberger is an actuary working as a Senior Manager in Actuarial & Insurance Services at Deloitte, with a focus on GenAI applications in the insurance industry. During his previous tenure in the actuarial department at Allianz Private Health, he was responsible for various data- and AI-related topics for several years. He holds an MSc degree in Financial Mathematics and Actuarial Sciences from the Technical University of Munich, as well as an MSc degree in Computer Science and a PhD in Mathematical Stochastics from the University of Passau. Additionally, he is a member of the German Association of Actuaries, a Certified Actuarial Data Scientist, and a Certified Enterprise Risk Actuary. He is actively involved in several Actuarial Data Science committees of the German Association of Actuaries, serves as a workstream lead in the Artificial Intelligence Task Force of the International Actuarial Association, and is a member of EIOPA's Consultative Expert Group on Data Use in Insurance.