european actuarial academy

The European knowledge centre for actuaries

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

Simon Hatzesberger and Iris Nonneman

EAA's Anniversary Conference | 4 December 2025 | online

european actuarial academy

Presenters

Simon Hatzesberger

Deloitte

shatzesberger@deloitte.de

Iris Nonneman

Achmea

irisnonneman@gmail.com

Simon Hatzesberger is a certified actuary, Certified Actuarial Data Scientist (CADS), and Certified Enterprise Risk Actuary (CERA), currently serving as a Manager in Actuarial & Insurance Services at Deloitte. He actively participates in various national committees and working groups on data science and AI within the German Association of Actuaries (DAV). Internationally, he engages in the AI-DS Working Group of the Actuarial Association of Europe (AAE), is a member of EIOPA’s Consultative Expert Group on Data Use in Insurance and leads a workstream within the AI Task Force of the International Association of Actuaries (IAA).

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.

Agenda

- History and Introduction of Advanced AI
- Four Practical GenAI Case Studies
  - Case Study: Improving Claim Cost Prediction with LLM-Features
  - Case Study: GenAI-Driven Market Comparison
  - Case Study: Car Damage Classification with Vision-Enabled LLMs
  - Case Study: Multi-Agent System for Autonomous Data Analysis
- Further GenAI Applications in the Insurance Industry
- Q&A

european actuarial academy

History and Introduction of Advanced AI

History and Introduction of Advanced AI

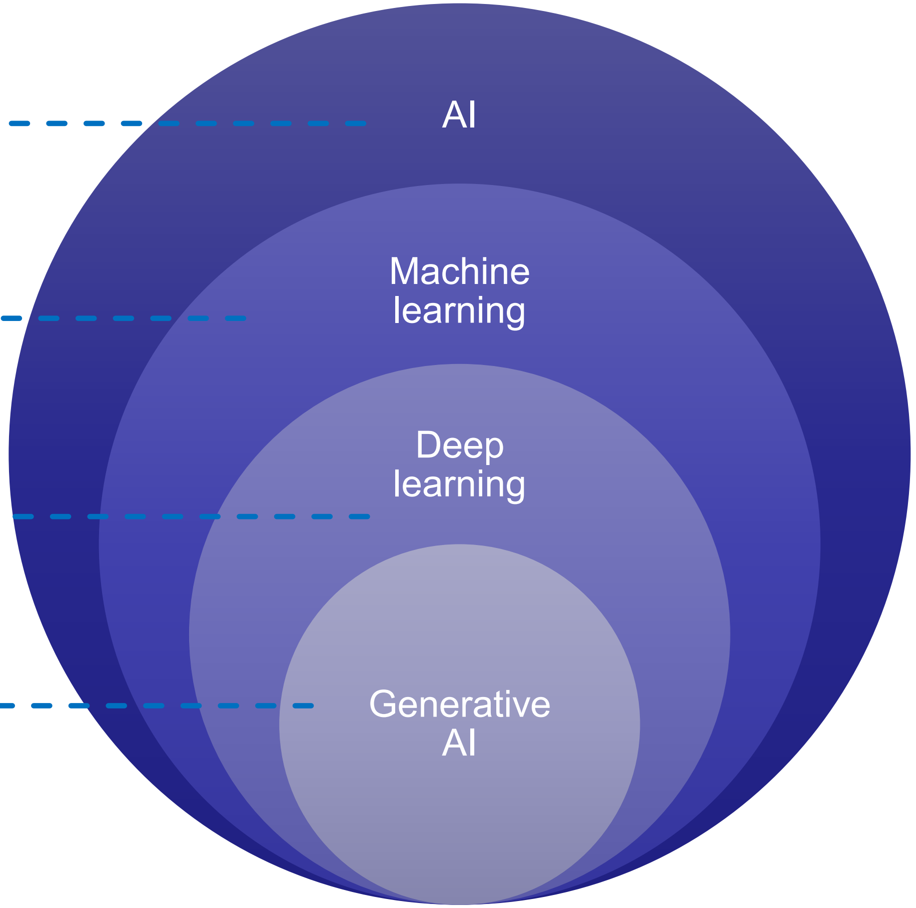
Artificial intelligence system (AI system) is a machine-based system that is designed to operate with varying levels of *autonomy* and that can, for explicit or implicit objectives, *generate output* such as predictions, recommendations, or decisions influencing physical or virtual environments (OECD).

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

Machine learning: subfield of AI, creates 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 and Introduction of Advanced AI

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

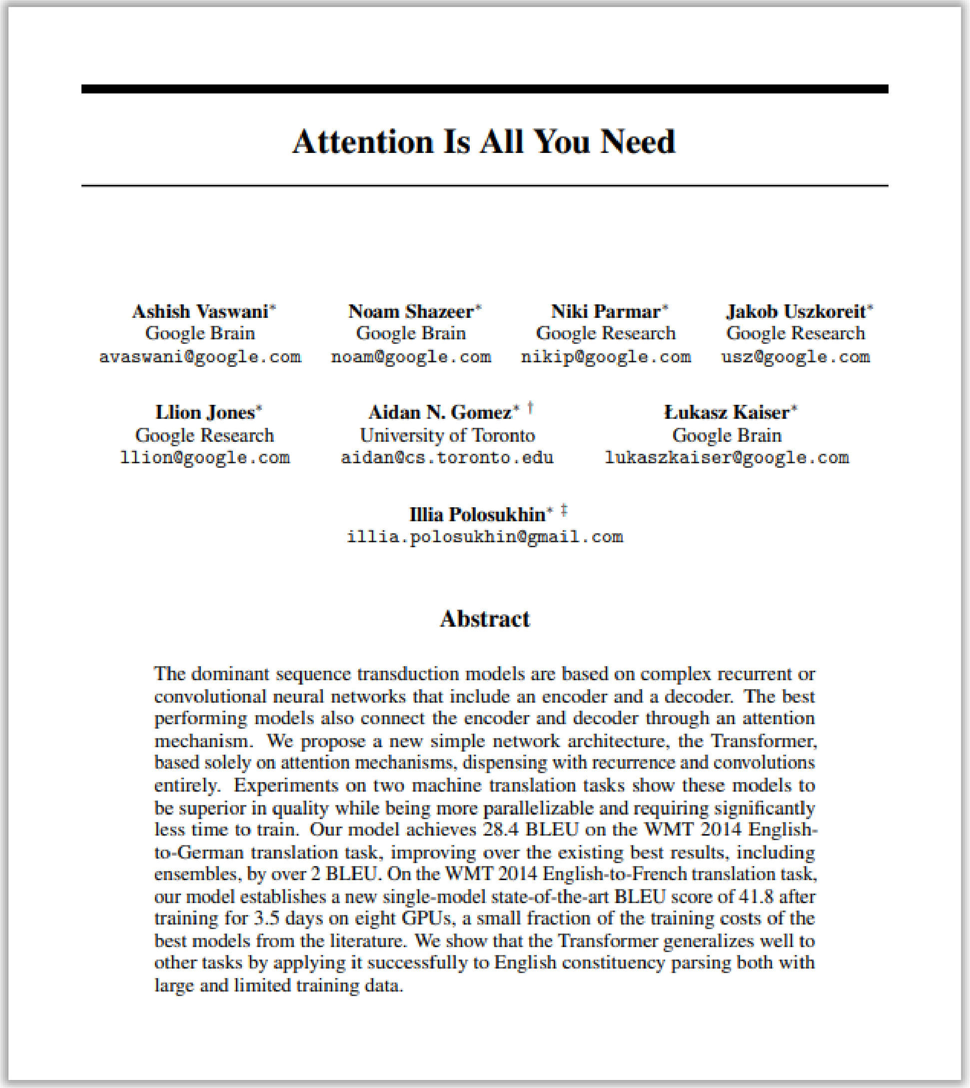
marked rapid progress with deep learning, AlexNet, and transformer models.

Advances of computing power and data availability

2017

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

Adoption of NNs and GenAI



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.

The rise of pattern recognition

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

Four Practical GenAI Case Studies

GenAI's Transformative Impact on Actuarial Science

Unlock Hidden Value: LLMs transform unstructured text into actionable insights

Automated Complex Workflows: RAG systems process lengthy documents in minutes vs. hours of manual work

Rich Contextual Understanding: Vision-enabled models (can) outperform traditional approaches while providing location and severity details

Scalable Agent Systems: Multi-agent frameworks enable autonomous analysis with modular, upgradeable components

Improving Claim Cost Prediction with LLM-Extracted 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

Constructing features

```
system_prompt_adjusted = '''
Your task is to extract structured information about injuries and cause of injury from the given text.
Follow this schema strictly:

- number_of_body_parts_injured: The total count of injured body parts.
- main_body_part_injured: The primary body part affected, described concisely (e.g., "HEAD", "THUMB").
- cause_of_injury: Specify by verb.
    - If verb given: return only the primary action verb that directly caused the injury (e.g. "fall" not "fell from box")
    - If cause is not mentioned, infer from context if possible, otherwise return "unspecified".

Ensure accuracy and consistency in the extracted details. Do not add interpretations beyond the provided text.
'''
```

Case Study: Improving Claim Cost Prediction w/ LLM-Features

Constructing features

	Log_UltimateIncurredClaimCost	WeeklyWages	Gender_M	PartTimeFullTime_F	DaysWorkedPerWeek	body_part_category	cause_of_injury_category	number_of_body_parts_injured
5	10.749461	651.0	1	1	5	HAND_FINGERS	LACERATION	1
6	8.791520	205.0	1	0	3	LOWER_EXTREMITY	FALL_SLIP_TRIP	1
7	9.659749	200.0	0	0	2	TORSO	FALL_SLIP_TRIP	1
8	8.632953	440.0	1	1	4	TORSO	FALL_SLIP_TRIP	2
9	7.074824	200.0	0	1	5	HAND_FINGERS	LACERATION	1

Case Study: Improving Claim Cost Prediction w/ LLM-Features

Prompt to LLM

Input is claim textual data

LLM features

Output is three new columns with features

Categorize features

Write a description of the significance of this event

Predict claim costs with GBM

Add LLM features to the feature space with traditional numerical and categorical features

Results

Higher accuracy

More insights into cause of claim costs

Case Study: Improving Claim Cost Prediction w/ LLM-Features

Process and results

Textual data to prompt → LLM-features → GBM

- **Improved model performance and predictive LLM features**
- **Business Value**: Enhanced risk assessment and early identification of high-cost claims
- **Insight**: LLMs successfully transform unstructured text into powerful predictive features

european actuarial academy

Case Study: GenAI-Driven Market Comparison

Setting & Problem

Market comparisons involve extracting and analyzing 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.

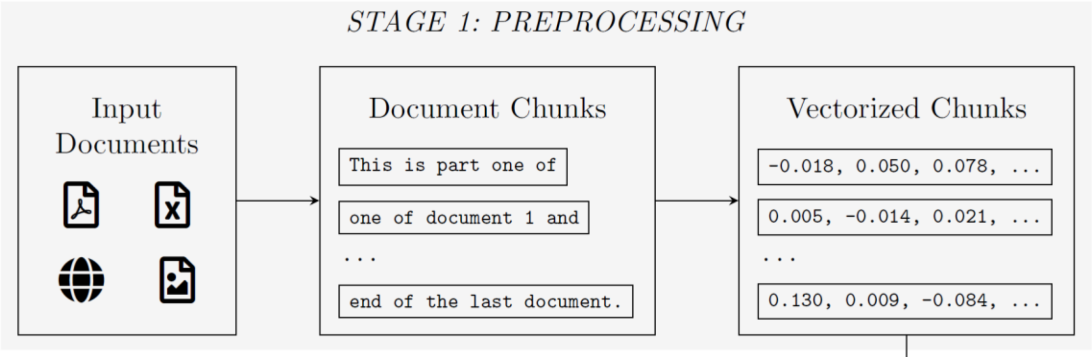
→ **Strategic Blind Spots**
Inefficient processes delay decision-making and limit competitive insight.

Solution

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

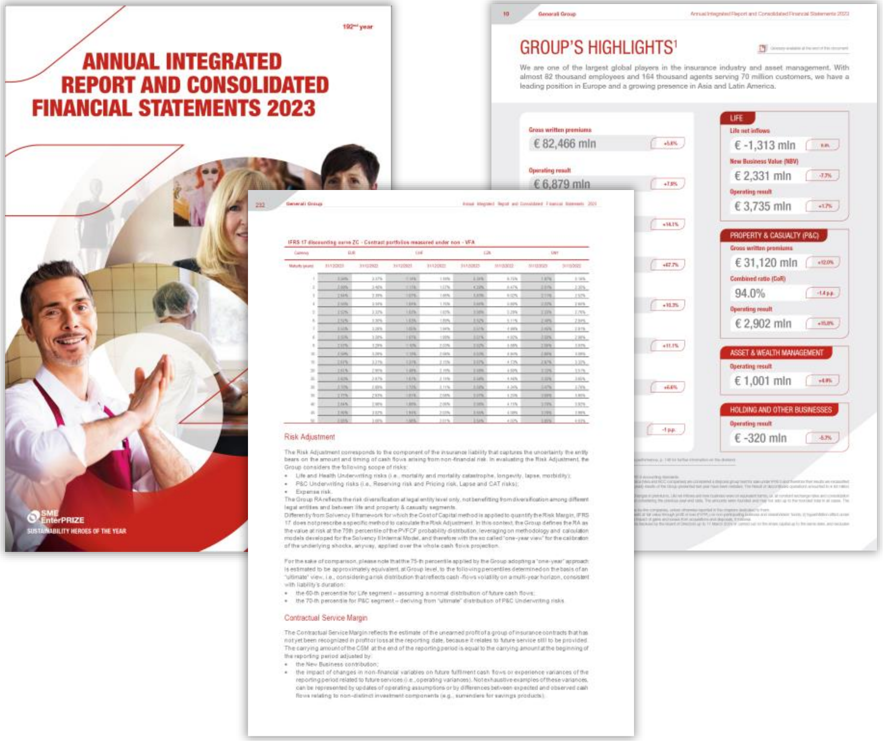
→ **Cross-Domain Applicability**
Solution applies to annual reports, tariff comparisons, etc.

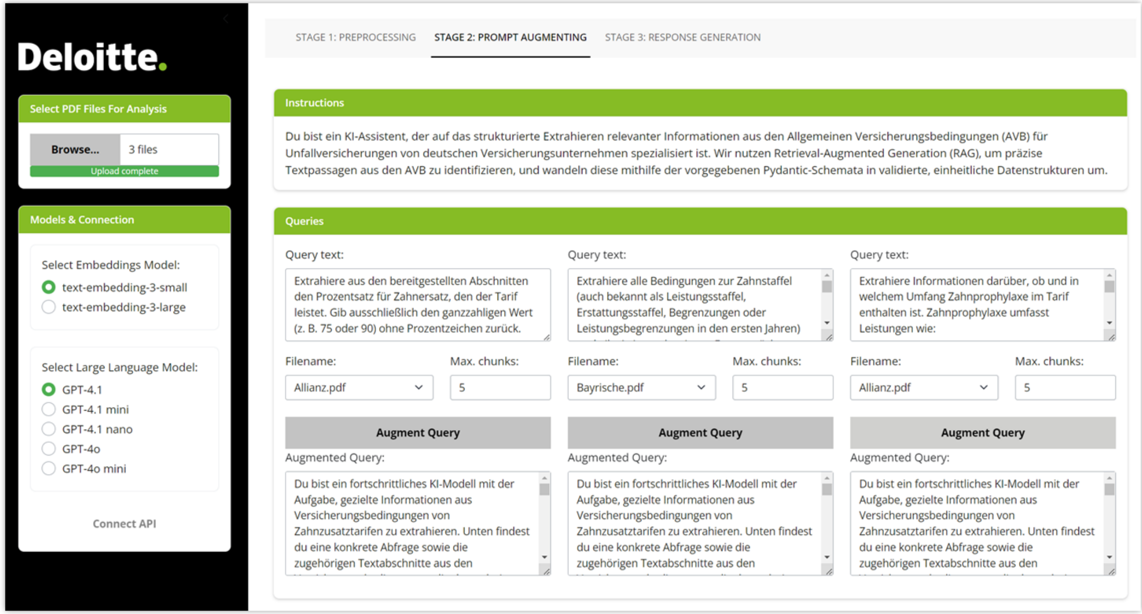
3-Stage RAG Pipeline for Structured Data Extraction



1 / **Preprocessing**
Convert, clean, and chunk documents; encode chunks into embeddings for efficient similarity search.

Case Study: GenAI-Driven Market Comparison





Annual Reports of
Generali, Allianz & Zurich

Live Demonstration of
GenAI-Driven Market Comparison

Case Study: Car Damage Classification w/ Vision-Enabled LLMs

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 labeled car damage to classify images 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.

Case Study: Car Damage Classification w/ Vision-Enabled LLMs

Results

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

Finding 1:

→ Non-fine-tuned vision-enabled LLMs already perform well out of the box

→ Fine-tuning boosts classification performance

Finding 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

european actuarial academy

Case Study: Multi-Agent System for Autonomous Data Analysis

Setting & Problem

Actuarial and data teams regularly receive new tabular datasets and **need quick, consistent exploratory analysis and reporting** without manually writing code each time.

→ Exploratory data analysis and report writing are still **largely manual**, **time-consuming**, and **hard to standardize** across datasets and teams.

→ Single LLM prompts **struggle to reliably handle multi-step workflows** (EDA, visualization, web lookup, report drafting) in a robust way.

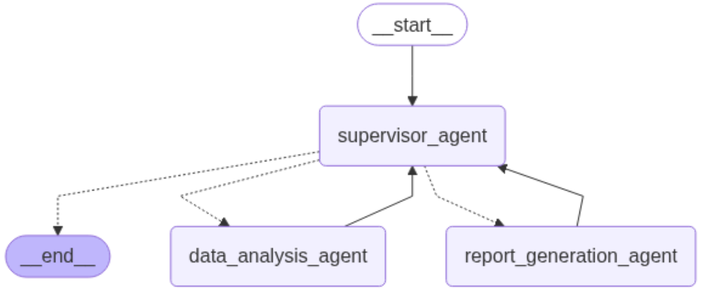
Solution

→ **Multi-Agent Systems**
Build a multi-agent system with specialized data-analysis, report-generation, and supervisor agents that collaboratively run EDA on any tabular dataset.

→ Let agents **call tools** (code interpreter, plotting, web search) and **exchange structured outputs** to automatically produce a consistent, well-structured analysis report **with minimal human intervention**.

Case Study: Multi-Agent System for Autonomous Data Analysis

Architecture



Data Analysis Agent *(GPT-4.1)*: EDA, statistics, visualizations

Report Generation Agent *(o1)*: Narrative synthesis, web research for context

Supervisor Agent *(GPT-4.1 mini)*: Workflow coordination

LangGraph as the Orchestration Framework

Results

Introduction

This dataset, commonly referred to as the "Diabetes 130-US hospitals" dataset, comes from a large pool of patient admissions spanning the years 1999–2008 at 130 U.S. hospitals. Each record represents a hospital stay for a diabetic patient, including demographic information, diagnoses, length of stay, lab results, and medications used. This dataset is valuable for analyzing factors associated with readmissions, medication changes, and general outcomes in diabetic care.

Column Descriptions

Below is an overview of each column's meaning and context:

- **encounter_id** – Unique identifier for the encounter.
- **patient_nbr** – Unique identifier for the patient (multiple encounters possible).
- **race** – Patient-reported race (Caucasian, AfricanAmerican, etc.).
- **gender** – Gender of the patient (Male, Female, or Unknown).
- **age** – Patient's age in 10-year intervals (e.g., [0-10), [10-20), [90-100)).
- **weight** – Weight recorded in specific ranges (mostly missing).
- **admission_type_id** – Numeric ID describing the admission category (e.g., emergency, urgent).
- **discharge_disposition_id** – Numeric ID describing discharge status (home, transfer, etc.).
- **admission_source_id** – Numeric ID describing the referral source (e.g., physician referral, emergency).
- **time_in_hospital** – Number of days between admission and discharge (1–14).
- **payer_code** – Payment source (e.g., Medicare, Medicaid, private insurance).
- **medical_specialty** – Primary specialty of the attending physician (e.g., InternalMedicine).
- **num_lab_procedures** – Number of lab tests performed during the encounter.
- **num_procedures** – Number of non-lab procedures performed.
- **num_medications** – Number of distinct medications administered.
- **number_outpatient** – Count of prior outpatient visits within one year.
- **number_emergency** – Count of prior emergency visits within one year.
- **number_inpatient** – Count of prior inpatient visits within one year.
- **diag_1** – Primary diagnosis (ICD-9 code).
- **diag_2** – Secondary diagnosis (ICD-9 code).
- **diag_3** – Tertiary diagnosis (ICD-9 code).
- **number_diagnoses** – Number of diagnoses on record for the encounter.
- **max_glu_serum** – Indicates glucose serum measurement range (>200, >300, Normal, or none).
- **A1Cresult** – Shows A1C test results (>7, >8, Normal, or none).
- **metformin, repaglinide, ..., insulin** – Prescription status (No, Steady, Up, Down) of various diabetic medications.
- **change** – Indicates whether the diabetes medication dosage was changed during the encounter.
- **diabetesMed** – Indicates whether any diabetes medication was prescribed during the encounter.
- **readmitted** – Readmission status: "<30" (within 30 days), ">30" (after 30 days), or "NO" (no readmission).

Data Preview

Below are the first 10 rows from the dataset:

Missing Values Summary

Below is the count of missing values among selected columns:

- **max_glu_serum**: 96,420 missing
- **A1Cresult**: 84,748 missing
- **weight**: 98,569 missing

Most essential demographic and numeric columns have complete data. However, lab-related fields (max_glu_serum, A1Cresult) and weight are frequently missing, which could impact models needing these predictors.

Descriptive Statistics

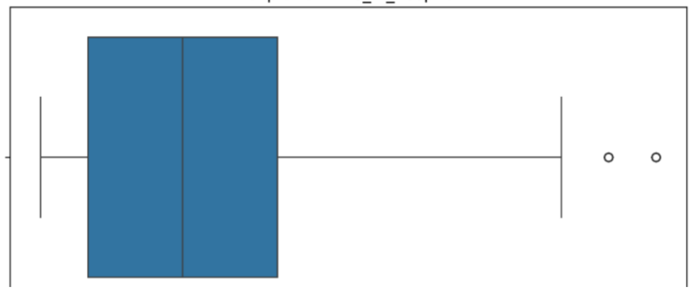
Numerical Columns

Below is a summary of the numeric features:

Column	Count	Mean	Std Dev	Min	25%	50%	75%	Max
time_in_hospital	101,766	4.40	2.99	1	2	4	6	14
num_lab_procedures	101,766	43.10	19.67	1	31	44	57	132
num_procedures	101,766	1.34	1.70	0	0	1	2	6
num_medications	101,766	16.02	8.13	1	10	15	20	81
number_outpatient	101,766	0.37	1.27	0	0	0	0	42
number_emergency	101,766	0.20	0.93	0	0	0	0	76
number_inpatient	101,766	0.64	1.26	0	0	0	1	21
number_diagnoses	101,766	7.42	1.93	1	6	8	9	16

Immediately below are the boxplots for each of these numerical columns, which help visualize their distributions and potential outliers:

Boxplot of time_in_hospital



Further GenAI Applications in the Insurance Industry

Further GenAI Applications in the Insurance Industry

Automated Reporting

Use prompt engineering + fine-tuned models to automate accurate, consis-tent regulatory reports (e.g., SFCRs).

Customer Interaction and Support

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

Claims Processing

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

Modernizing Legacy Systems

Use Generative AI to translate, refactor legacy code into modern languages and frameworks, easing migration.

Employee Training

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

...

Q&A

european actuarial academy

The European knowledge centre for actuaries

Thank you!

Please rate the conference via the survey-link you will receive per email.

Visit our website

www.actuarial-academy.com

for more events.

Follow us on LinkedIn

www.linkedin.com/company/642904

for updates & actuarial fun.