Democratizing Actuarial Expertise Through Fine-Tuned Chain of Thoughts

Manuel Caccone & Claudio Senatore Reso

Quant Actuaries | Italian Society of Actuaries & AAE, AI Task Force

6th European
Congress of Actuaries
www.eca2026.org

01

Intro

How do we democratize actuarial knowledge without compromising rigor?

Limited Access

Actuarial expertise confined to specialists. Small insurers can't afford large teams.

AI Integration

Leverage LLMs to disseminate actuarial knowledge at scale.

Maintain Rigor

Fine-tuned "Chain of Thoughts" preserves professional accuracy and auditability.

Democratized Access

Business analysts get professional-grade actuarial analysis instantly.

02

The Problem & Our Solution

The Problem & Our Solution

The Challenge

Insurance products are complex. Actuarial expertise is scarce.

The Question

What if analysts could ask actuarial questions and get professional analysis instantly?

The Solution

LLMs + "Actuarial Chain of Thoughts" fine-tuned on authoritative literature.

Language Support

8 languages — covers 90%+ of actuaries worldwide. Native thinking, not translation.

03

Three-Model Architecture & LoRA Technology

Why LoRA? — Efficient Fine-Tuning

LoRA: instead of retraining the whole model, LoRA adds small "plug-in" layers, far cheaper, and swappable per task.

Faster & Lighter

Far less memory and compute than full fine-tuning

Modular

Swap adapters in or out — no base retraining

We built three adapters, each with one job:

Simplifier

Turns business questions into actuarial reasoning

Reporter

Writes clear, formatted reports

Excelifier

Builds interactive Excel workbooks

Base Model & Why LoRA — Technical Choices

Llama 3.1 8B Instruct

Instruction-tuned baseline · 8B parameters, fits in a single 24 GB GPU · 128,000-token context window · Apache 2.0 licence · strong multilingual performance across the 8 target languages

Why 8B? — Model Size Trade-off

8B is the "sweet spot": enough capacity for actuarial chain-of-thought reasoning, trainable on one GPU without a cluster. Models in the 7–14B range (QWEN 7B/14B, Mistral 7B) offer similar capability — Llama 3.1 was preferred for its instruction-following quality on structured, multi-step reasoning tasks.

Why LoRA — Not Full Fine-Tuning

Full fine-tuning of an 8B model requires ~80 GB of VRAM, weeks of GPU time, and risks catastrophic forgetting of the base knowledge. LoRA trains only small adapter matrices injected into the attention layers (<1% of total parameters): ~24 GB VRAM, hours of training, base weights remain frozen and untouched.

LoRA Keeps the Base Intact

The 8B base model is loaded once and never modified. Each LoRA adapter is a ~50 MB file hot-swapped at inference time. Three tasks, three independent adapters — zero interference between them, zero retraining of the base for each new specialisation.

04

Training Data & Design

TRAINING DATA The training corpus is not input data. It is the definition of what the model becomes.

A fine-tuned model learns the specific reasoning patterns in its training examples, no more, no less. Content selection is not a data-preparation step: it is the primary design decision.

The choice here was operational specificity across the full breadth of actuarial practice. Three domains below are illustrative, not a complete taxonomy:

SYSTEM You are an expert actuary…

USER What is our reserve adequacy…

ASSISTANT [Chain-of-thought reasoning]

Loss computed only on ASSISTANT tokens — model trains on responses, not questions.

Illustrative domains:

Domain	Representative task
Loss reserving	IBNR chain-of-thought, run-off triangles, adequacy assessment
Premium calculation	Frequency–severity decomposition, rate change derivation, technical pricing
Technical bases	Survival function analysis, mortality derivation, credibility weighting

** Three illustrative examples — corpus spans the full breadth of actuarial practice.*

> 50,000 actuarial pages studied

Training corpus drawn from authoritative literature across all major actuarial domains

Caccone, M. & Senatore Reso, C. (2026). Fine-Tuning LLMs for Actuarial Expertise. Dataset: manuelcaccone/actuarial-gpt-conversations, Hugging Face Hub.

Training operates on two fronts: the questions you write, and the answers you pair with them.

Question axis: vocabulary level, technical register, detail in the prompt. Answer axis: depth, precision, structure — calibrated relative to the question.

Strategy	Question	Answer	Behavior
Symmetric	High detail	High detail	Expert-to-expert
Graduated	Variable	Calibrated	Adapts to input
Asymmetric uplift	Vague / low	High precision	Removes access barrier

① **Example Registry** — Domain tag, complexity, depth, review status — no example enters training without confirmed review.

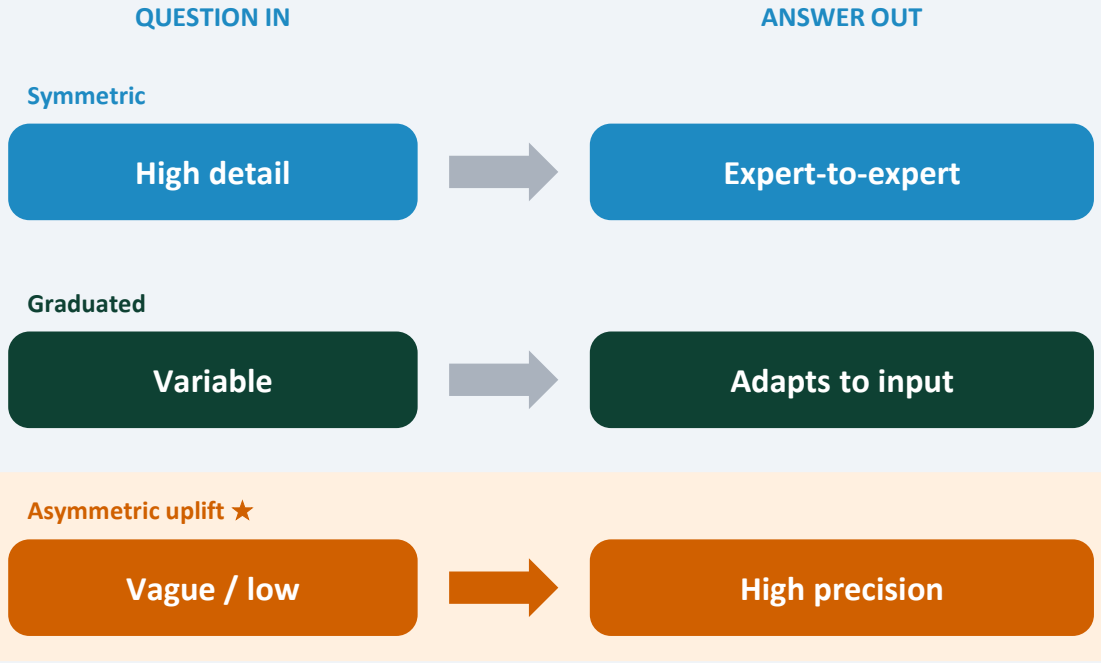
② **Reasoning Quality Gate** — Realistic problem? Correct reasoning depth? Verifiable output? Failure → revision, not bypass.

③ **Version-Controlled Runs** — Each run linked to an approved snapshot. Degradation traceable to the specific addition.

Caccone, M. & Senatore Reso, C. (2026). Fine-Tuning LLMs for Actuarial Expertise. Unpublished manuscript.

From question to answer

Same model — three ways to pair questions with answers



★Our choice: even a vague question returns a precise, professional answer — removing the access barrier.

What has been established. What follows.

01

The corpus is the model

Content selection, domain coverage, and the operational specificity of every example determine what the model can and cannot do.

Gaps in training data are gaps in professional competence.

02

Questions and answers are both design parameters

The vocabulary level of the prompt and the depth of the response are independent choices.

Aligning them — or deliberately misaligning them — shapes the professional behavior of the model.

03

Control is not optional

A platform that makes each training example reviewable, classifiable, and version-controlled is not infrastructure overhead.

It is the mechanism by which training decisions remain auditable.

What follows: architecture, training methodology, measured outcomes, and open questions.

05

Production Reliability & User Experience

Agentic Self-Correction Framework

Problem: AI-generated code can fail without anyone noticing.

Solution: an automatic loop that catches and fixes its own errors.

– Run the code, then detect and sort any failures automatically

– Fix the errors and re-run to confirm the code works — no human needed

06

Economics & Open Source Philosophy

…this is where democratization becomes real.

Traditional ML

High infrastructure costs
Vendor lock-in
Expensive full retraining

Serverless Architecture

Pay-per-use
Auto-scaling
Zero idle cost

Open Source

Community-driven
Fully customizable
No proprietary dependency

Vendor Independence

Deploy anywhere
Switch providers freely
Docker containerized

07

Real-World Impact — Case Studies

Actuarial Team Capacity — Before vs. After Automation

85%

time reduction
in pricing cycle

Before Automation:

– 40% of team capacity on routine tasks

– 20 hours per line of business per cycle

After Automation:

– Full analysis completed in ~3 hours total

– Team freed for strategic & creative work

Areas: Pricing · Reserving · Risk Mgmt · Competitive Analysis

08

Transparency, Compliance & Key Benefits

How to enhance actuarial work through AI?

Economics

Significant cost reduction vs. traditional methods.

Transparency

Complete explainability & audit trails for compliance.

Speed

Full analyses completed in seconds.

Democratization

Sophisticated analysis accessible to non-specialists.

Prof. Elevation

Actuaries focus on strategic & creative tasks.

Chain of Thoughts

Reasoning explicit, auditable, and trustworthy.

09

Technical Innovation Summary

Technical Innovation Summary

Novel Training — 15,000 curated examples from authoritative actuarial literature

Specialized Architecture — 3 focused models, each optimized for its task

Agentic Validation — automatic error detection & self-correction

Multilingual — 8 languages, 90%+ of global actuaries covered

Chain of Thoughts — explicit, auditable, trustworthy reasoning

LoRA Efficiency — smaller models, faster training, lower cost, modular

Serverless Deployment — pay-per-use, auto-scaling, no vendor lock-in

10

Limitations, Future Work & Conclusions

Limitations, Future Work & Conclusions

System Limitations

Constrained to training corpus domains.
Hallucination risk on novel edge cases.

Model Refinement

Expand corpus to new domains & regulatory frameworks.
Real-time data integration.

Community

Open-source release.
Dataset on Hugging Face Hub.
Contributions welcome.

Conclusions

Train AI on actuarial literature.
Democratize access.
Elevate the profession.

Everything is public — contribute, reproduce, build on it.

Dataset	manuelcaccone/actuarial-gpt-conversations · Hugging Face Hub 15,000+ curated Q-A pairs across all actuarial domains
Model Weights	Fine-tuned LoRA adapters: Simplifier, Reporter, Excelifier Published on Hugging Face Hub
Code & Framework	Full training pipeline · agentic validation loop · serverless deployment GitHub: https://github.com/manuelcaccone/actuarial-intelligence
License	Apache 2.0 — free for research, commercial use, and modification Contributions from the actuarial community are welcome

Scan to access the repo

https://github.com/manuelcaccone/actuarial-intelligence

We always overestimate the change that will occur in the next two years
and underestimate the change that will occur in the next ten.

Bill Gates

Thank You!

- Do you have any questions?
- Everything will be published — stay tuned!

- Manuel Caccone
- Quant Actuary | AI Engineer | Risk Manager | Data Scientist

- Claudio Senatore Reso
- Actuary | SAS Institute Italian | Vice – Chair AI Task Force EAA

- Emails: manuel.caccone@gmail.com – claudio.senatore@gmail.com
- Phone: +39 320 366 5598
- GitHub / HuggingFace: manuelcaccone - SenCla