AROCHA & ASSOCIATES

Practical Integration of Feed-Forward Neural Networks into Core Actuarial Workflows

INSURANCE DATA SCIENCE CONFERENCE 2026
HOUSE OF INSURANCE, LEBNIZ UNIVERSITÄT HANNOVER

Carlos Arocha, FSA, SCR
9 June 2026

AGENDA

01 Motivation

02 Case setup and actuarial framing

03 Baseline models: GLM Toolkit

04 FFNN challenger

05 Decision framework and adoption path

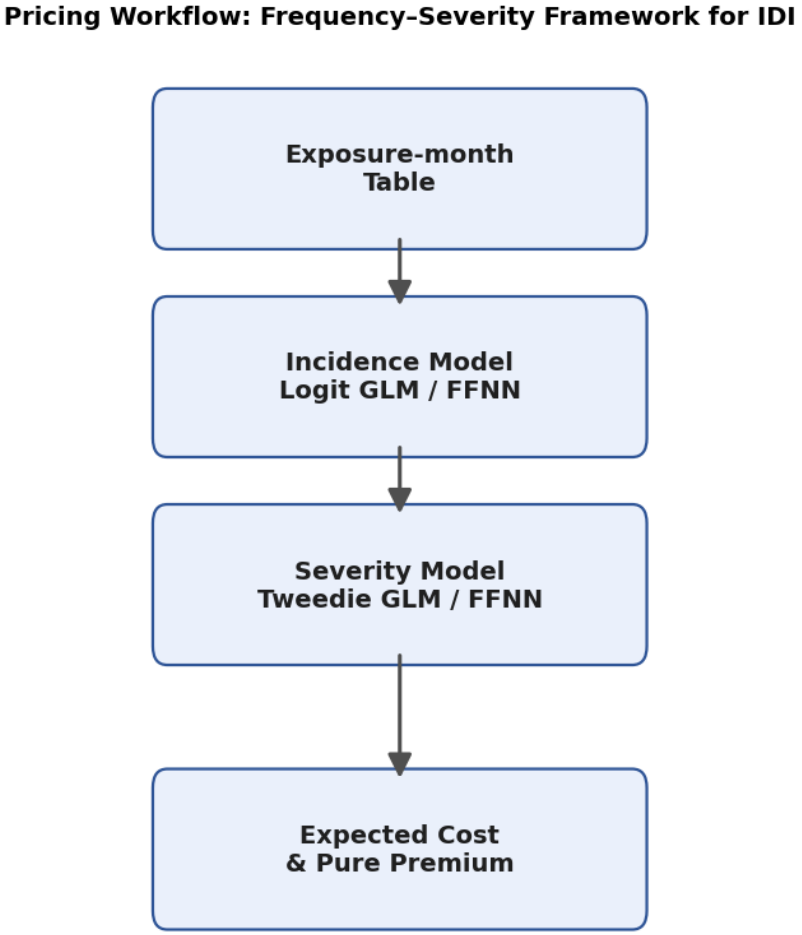
06 Conclusions

01 Motivation

GLMs vs Feed-Forward Neural Networks (FFNNs) in practice

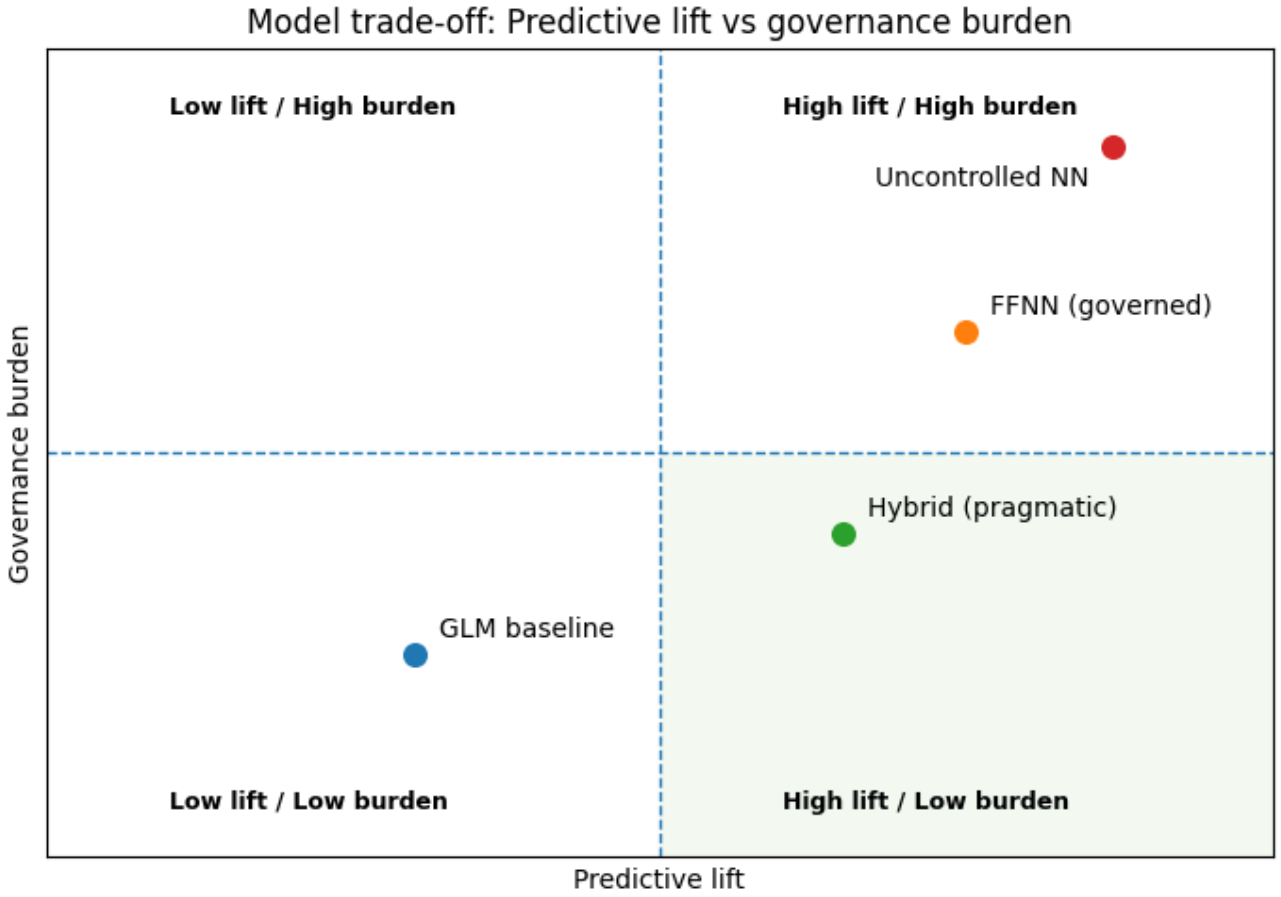
GLMs anchor governance; FFNNs add lift—best delivered via hybrids.

- Individual Disability Income (IDI) pricing case study using synthetic data
- Objective: incremental lift that remains governable
  - calibration
  - stability
  - audit trail
- Output is a deployable frequency—severity workflow



Why this matters now

AI is reshaping workflows: we need performance gains with audit-ready governance.



- Actuarial reality
  - accuracy alone is not sufficient; models must be explainable and controllable
- Typical blockers are
  - calibration risk
  - tail instability
  - weak drift controls
  - unclear documentation
- **Key question is:**
  - can FFNNs be adopted as structured extensions to GLMs rather than black-box replacements?

02 Case Setup and Actuarial Framing

The synthetic IDI dataset

Synthetic data lets us prove the end-to-end GLM-FFNN-Hybrid workflow safely.

- Two files:
  - exposure—month (onset)
  - claims (ultimate cost)
- Modeling challenges baked in:
  - Rare events
  - Heavy-tailed costs
  - High-cardinality factors (region / occupation / employer / plan)
  - Climate impacts
- A synthetic design enables open sharing and controllable experiments without confidentiality risk

Metric	Value
Exposure rows	144,000
Claim rows	1,565
Date min	2022-01-01
Date max	2023-12-31
Monthly onset rate	0.010868
Mean claim cost (CHF)	101,625
P90 claim cost (CHF)	160,396

Frequency—severity and pure premium

Pure premium (PP) is the governed bridge from models to price.

- Incidence: $p_i = \mathbb{P}[\text{onset}{=}1 \mid X_i]$
- Severity: $\mu_i = \mathbb{E}[\text{cost} \mid \text{onset} = 1, X_i]$
- Pricing target: $\mathbb{E}[\text{cost}] = \sum_i p_i \times \hat{\mu}_i$; pure premium per exposure-month

The layers will be evaluated using specific metrics

Layer	Metric	Purpose
Incidence	PR-AUC	Rare-event ranking quality
	Lift@10	Claims captured in top 10% risk
	Brier score	Probability accuracy (calibration-sensitive)
	Calibration	Reliability curve + mean p vs observed
Severity	MAE (CHF)	Overall error on claim costs
	RMSE log1p	Multiplicative error; tail-robust-ish
	Tail MAE (P90+)	Large-loss focus
Aggregate	PP error	Pure premium adequacy vs. actual
	O/E	Portfolio adequacy (Actual ÷ Expected)

Governance-first validation design

Design validation first, then models: calibrate, stress-test, and prove stability before benchmarking.

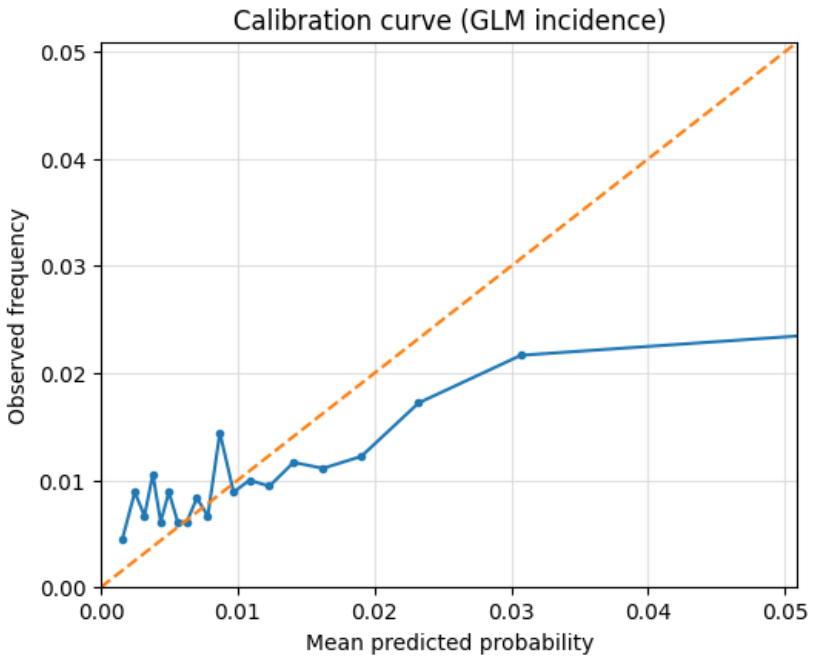
- Time-based split
  - Train on earlier months (24)
  - Test on most recent months (12)
- Same split + same features for all models to achieve a fair comparison
- Decision criteria emphasize:
  - Calibration
  - Stability
  - Tail behaviour
  - Adequacy

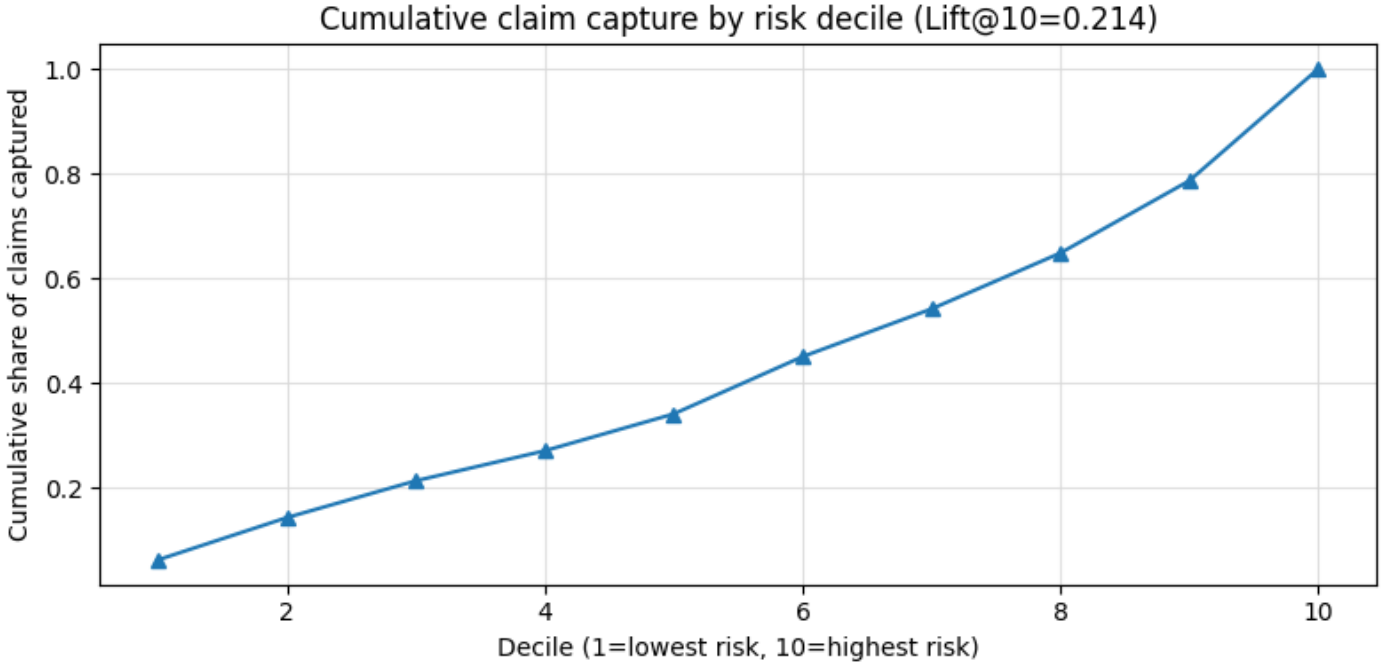
03 Baseline models: GLM Toolkit

Incidence baseline: Logit GLM analogue (sparse + regularised)

A sparse, regularised logit GLM typically gives a stable, audit-ready incidence baseline which is strong enough to benchmark any FFNN lift.

- Logistic regression on exposure-month rows (one-hot categoricals + scaled numerics)
- Rare event focus: PR-AUC and Lift@10 matter more than ROC AUCc

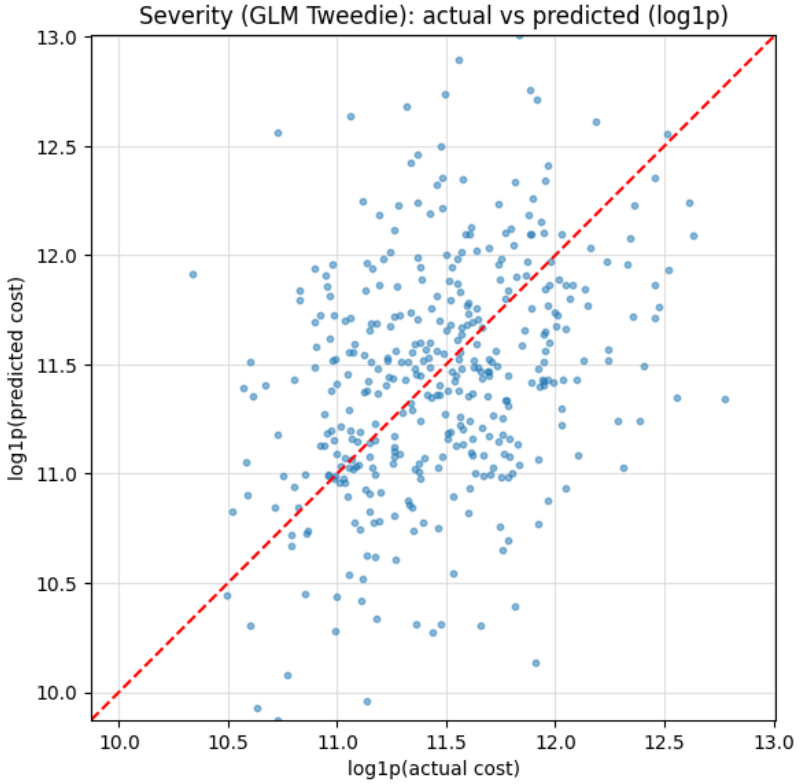


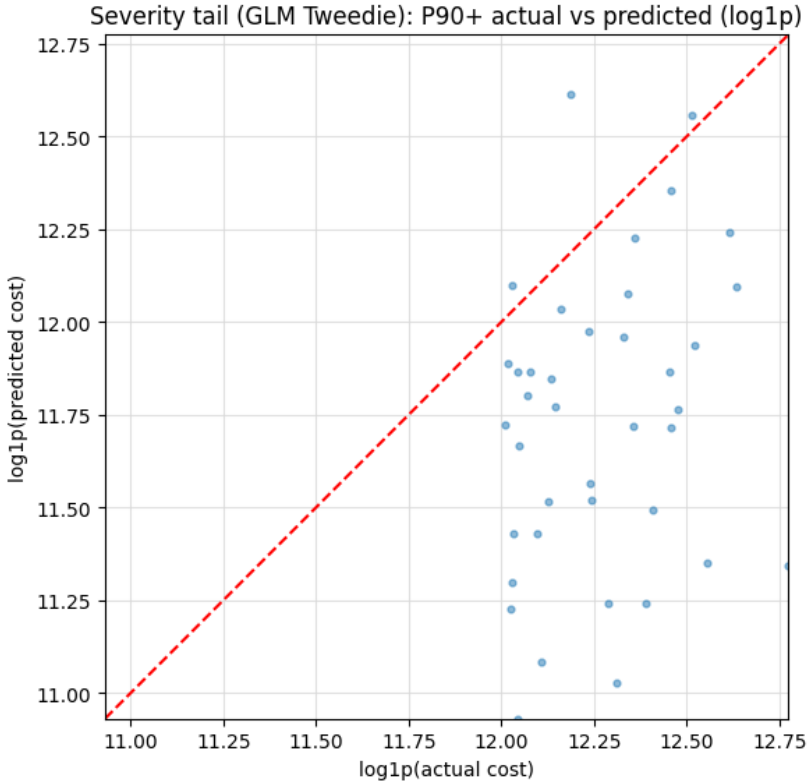


Severity baseline: Tweedie GLM (log link)

May be ideal as the audit-ready benchmark for any FFNN gain.

- Tweedie with log link is a robust baseline for skewed positive distributions
- Evaluate central accuracy and tail behaviour





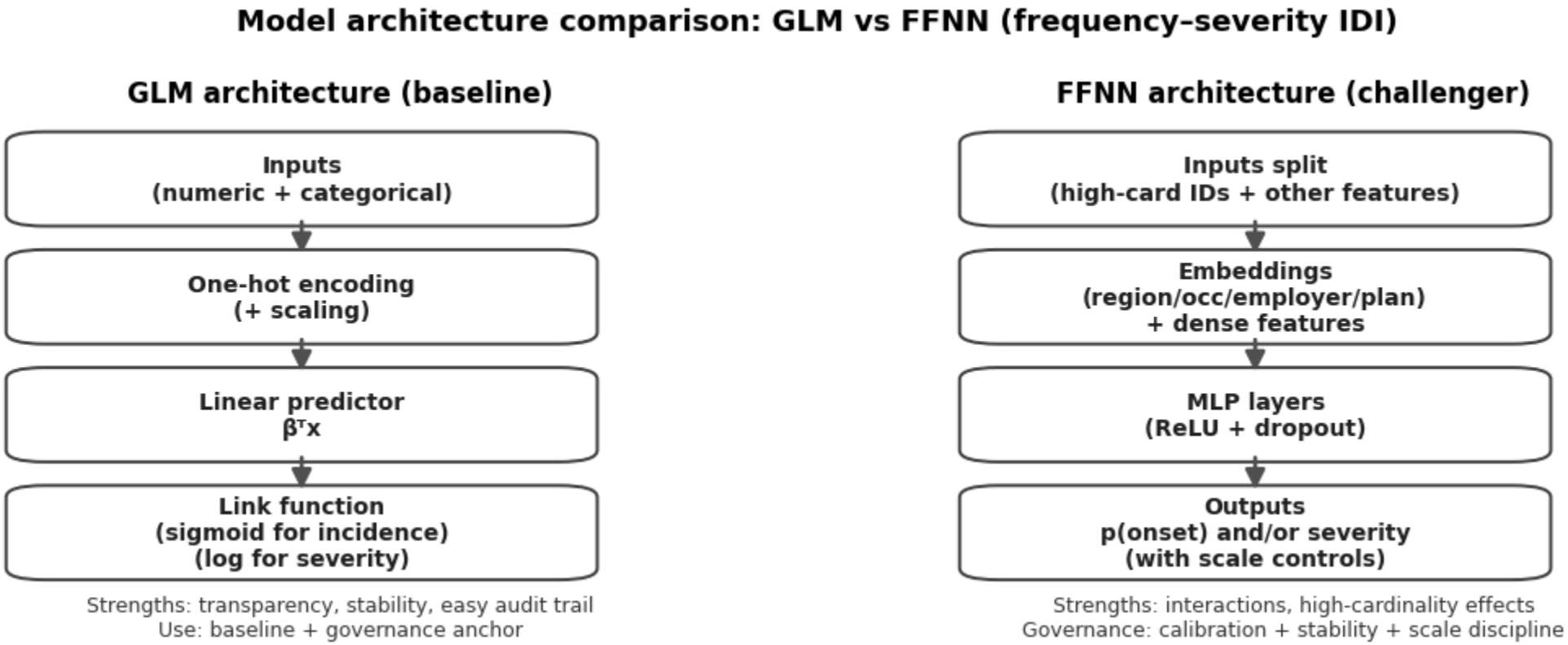
Mean actual (CHF)	106,135
Mean predicted (CHF)	106,932
MAE (CHF)	47,339
RMSE log1p	0.5587
Tail MAE P90+ (CHF)	89,130

04 FFNN Challenger

FFNN as a structured extension

FFNNs should avoid a “black box” connotation by proper governance and validation.

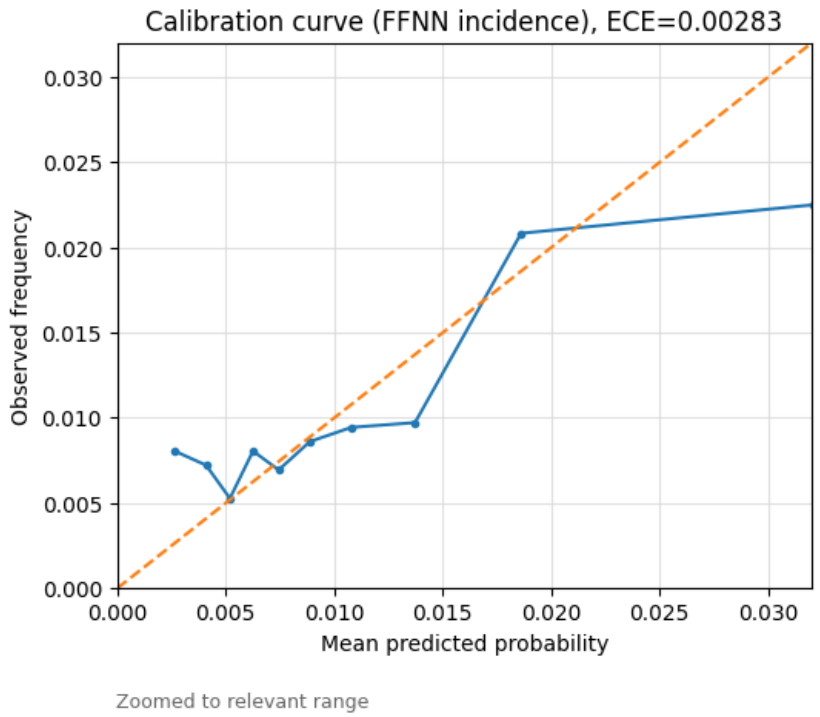
- Same frequency—severity actuarial structure
- Governance discipline: controlled transforms, bounded outputs, and reproducibility

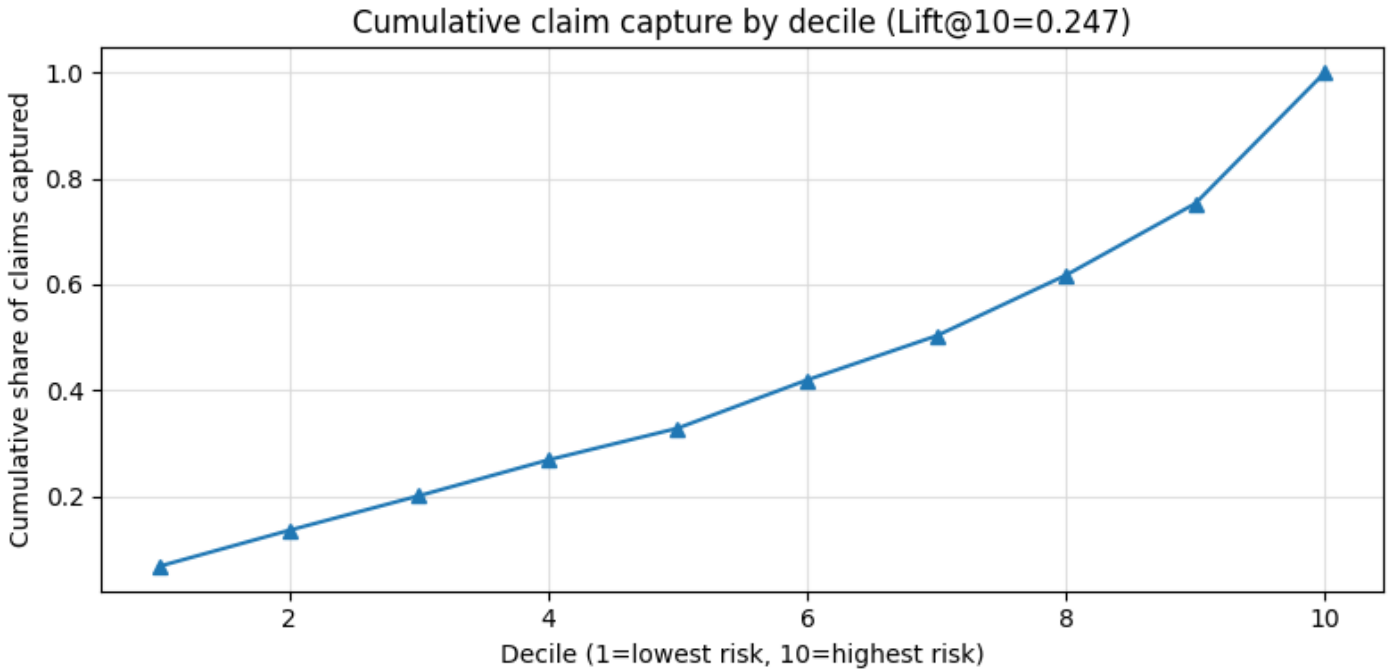


FFNN incidence

FFNN incidence improves rare-event ranking and calibration.

- ECE drops by 47% relative to GLM, i.e., the average absolute gap between predicted and observe incidence is substantially smaller
- Lift@10 increases 15%, i.e., top 10% highest-scored exposure-months capture a larger share of claims

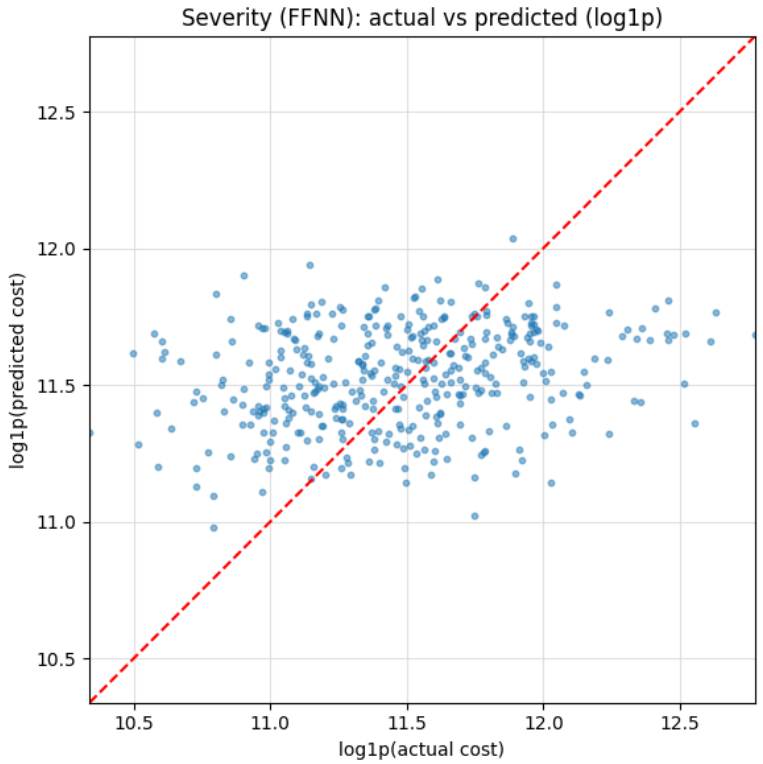


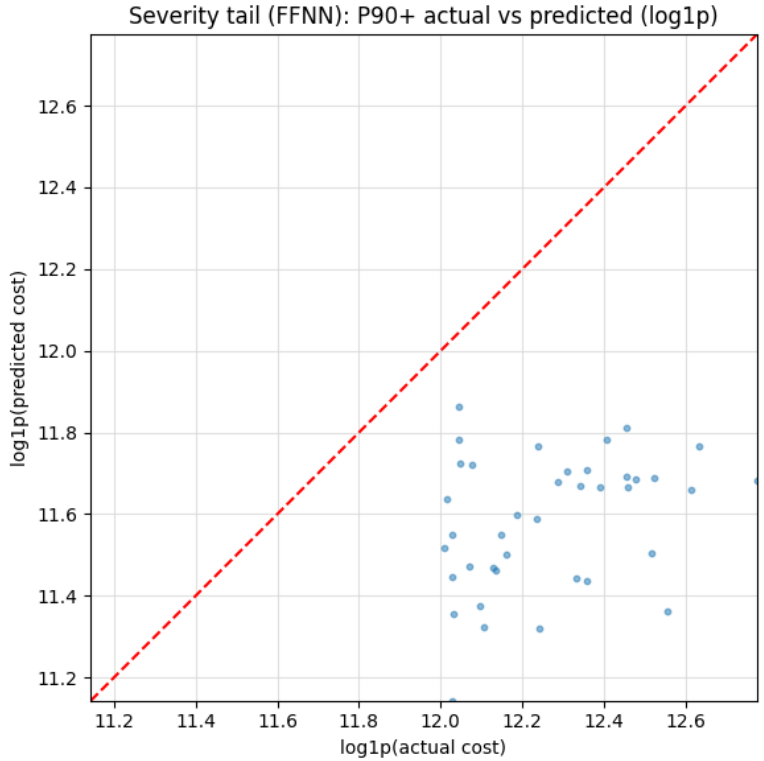


FFNN severity

FFNN severity improves fit on typical claims (MAE & RMSE) relative to GLM

- A lower RMSE suggests a better capture of the multiplicative structure of severity for the bulk of claims → smoother pricing relativities
- But a higher Tail MAE P90+ suggests less reliability where large claims concentrate





Mean actual (CHF)	106,135
Mean predicted (CHF)	103,042
MAE (CHF)	36,608
RMSE log1p	0.4313
Tail MAE P90+ (CHF)	108,032

Caveat emptor

Design choices: our team's governance principles

We do	We avoid	Governance rationale
Hybrid framework (GLM baseline + FFNN challenger)	Winner-takes-all replacement without a challenger	Preserves a stable benchnark; enables controlled promotion/rollback
Time split + rolling-window evaluation	Single random split as primary evidence	Reduces overfitting and improves stability evidence under drift
Rare-event metrics PR-AUC, Lift@10 + calibration checks	AUC-only conclusions	Pricing needs calibrated probabilities; ranking alone is insufficient
FFNN embeddings only for justified high-cardinality fields	Unbounded architecture growth / excessive depth	Controls complexity; improves reproducibility and reviewability
Severity trained on log1p(cost) with smearing + level calibration	Back-transform without bias correction (median-like predictions)	Avoids systematic underprediction on the CHF scale
Segment adequacy with credibility filters	Portfolio-only adequacy sign-off	Prevents hidden pockets of mispricing
Versioning + run metadata (seed, cutoff, library versions)	Results without reproducibility artifacts	Enables audit trail and repeatable validation

05 Decision Framework and adoption path

GLM vs FFNN

FFNN delivers better incidence calibration and risk ranking; GLM remains the stability anchor

Metric	Actual	GLM	FFNN	FFNN ÷ GLM
Observed rate	0.010667	0.012385 +16%	0.011223 +5%	0.9063
AUC (ROC)	-	0.6065	0.6380	
PR-AUC	-	0.0166	0.0192	
Brier	-	0.0106	0.0105	
Lift@10	-	0.2135	0.2578	
Mean cost (CHF)	106,135	106,932 +0.8%	102,292 −3.6%	
MAE (CHF)	-	47,339	36,534	
RMSE log1p	-	0.5587	0.4313	
Tail MAE P90+	-	89,130	108,032	
Pure premium	**1,132.11**	**1,310.62** **+16%**	**1,155.26** **+2%**	**0.8815**

FFNN relative to GLM:

- AUC uplift suggests a meaningful improvement in ranking power
- PR-AUC improvement is more meaningful for rare events because it focuses perramance in the “positive” class
- Brier change is negligible, which suggests both models have similar average probability error in the full population

Hybrid alternatives

Hybrid alternative isolate where the lift really comes from, so we can choose the strongest governance-to-performance trade-off.

Let

Total exposure = $36{,}000$ exposure-months

Actual cost $A = 40{,}755{,}967$

Expected cost $\hat{C} = \sum_i w_i \times \hat{p}_i \times \hat{\mu}_i$, where w_i is the exposure months weight, $\hat{p}_i$ is the predicted incidence probability and $\hat{\mu}_i$ is the predicted conditional mean claim cost

Then

$\widehat{PP}_A = \frac{\sum_i w_i \times \hat{p}_i^{FFNN} \times \hat{\mu}_i^{GLM}}{\sum_i w_i}$ is the hybrid A pure premium, and $\widehat{PP}_B = \frac{\sum_i w_i \times \hat{p}_i^{GLM} \times \hat{\mu}_i^{FFNN}}{\sum_i w_i}$

Model	Expected Total Cost (CHF)	Pure Premium (CHF / exp-mo)	O/E	PP Error vs Actual	Comment
GLM	47,182,238	1,310.62	0.864	+15.8%	Material overstatement
FFNN	40,963,136	1,137.86	0.995	+0.5%	Closest to adequacy
Hybrid A (FFNN inc, GLM sev)	41,837,421	1,162.15	0.974	+2.7%	Best hybrid bridge
Hybrid B (GLM inc, FFNN sev)	45,609,436	1,266.93	0.894	+11.9%	Still too conservative

Model comparison scorecard

A scorecard traces performance driver and supports a clear go/no-go decision.

Layer	Metric	Actual	GLM	FFNN	Hybrid A (FFNN inc × GLM sev)	Hybrid B (GLM inc × FFNN sev)	Notes
Incidence	Rate	0.010667	0.01239 +16.1%	0.01143 +7.2%	0.01143 +7.2%	0.01239 +16.1%	Calibration-in-the-large
	AUC (ROC)	-	0.60651	0.62469	0.62469	0.60651	Ranking quality
	PR-AUC	-	0.01658	0.01861	0.01861	0.01658	Rare-event ranking
	Brier	-	0.01063	0.01054	0.01054	0.01063	Probability accuracy
	Lift@10	-	0.21354	0.24740	0.24740	0.21354	Claims captured in top 10%
Claim cost	Mean cost (CHF)	106,135	106,932 +0.8%	103,042 −2.9%	106,932 +0.8%	103,042 −3.6%	Level check
	MAE (CHF)	-	47,339	36,608	47,339	36,608	Overall error
	RMSE log1p	-	0.55870	0.43131	0.55870	0.43131	Multiplicative error
	Tail MAE P90+	-	89,130	108,032	89,130	108,032	Large-loss focus
Aggregate	**Pure premium**	**1,132.11**	**1,310.62 +15.8%**	**1,181.92 +4.4%**	**1,200.23 +6.2%**	**1,280.65 +13.1%**	$p \times \mu$ over exposure

Pros and cons of the model alternatives

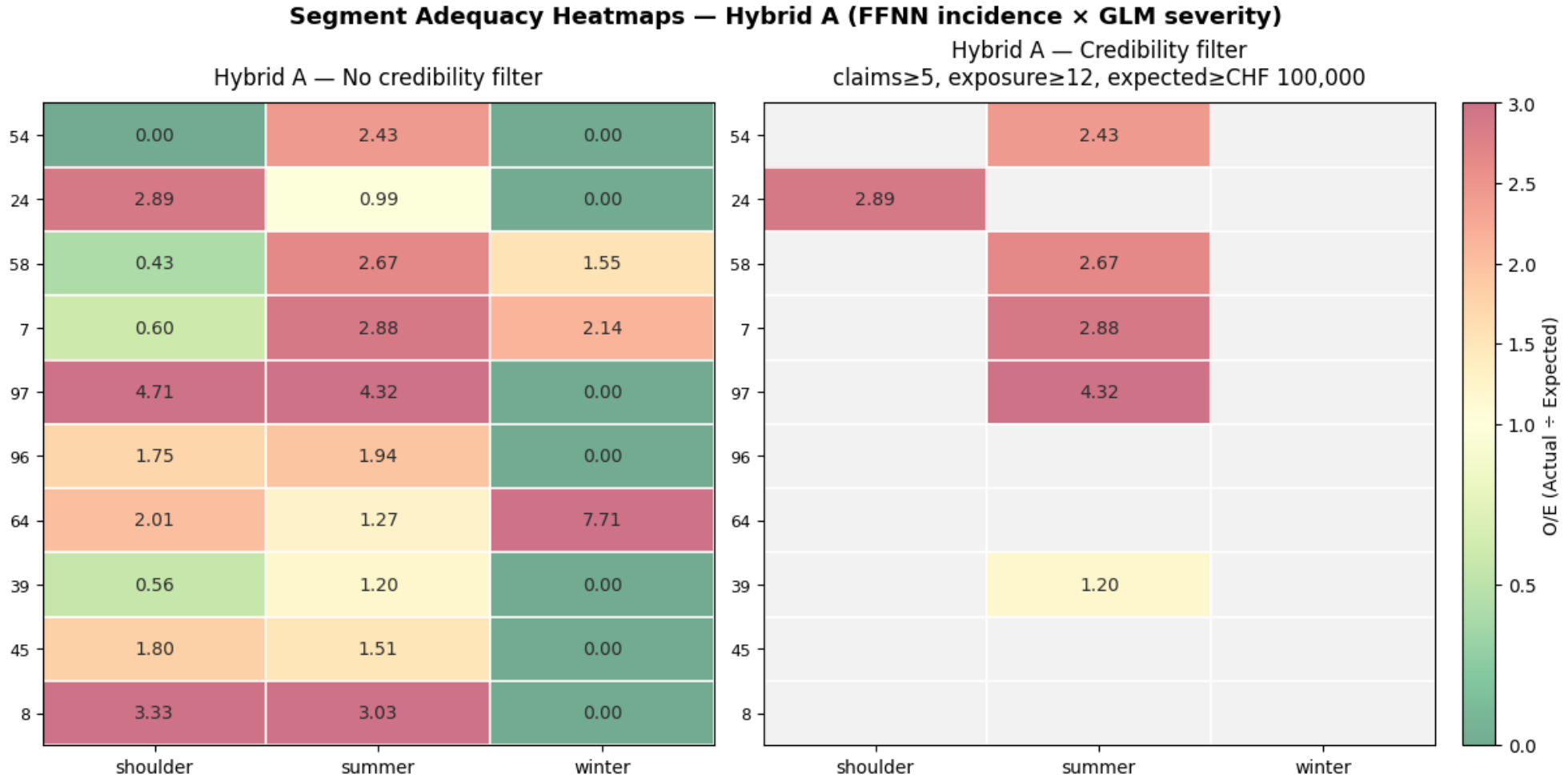
A pros- and cons- view translates performance into implementation trade-offs.

Model	Expected Total Cost (CHF)	Pure Premium (CHF / exp-mo)	O/E	PP Error vs Actual	Comment
GLM	47,182,238	1,310.62	0.864	+15.8%	Material overstatement
FFNN	40,963,136	1,137.86	0.995	+0.5%	Closest to adequacy
Hybrid A (FFNN inc × GLM sev)	41,837,421	1,162.15	0.974	+2.7%	Best hybrid bridge
Hybrid B (GLM inc × FFNN sev)	45,609,436	1,266.93	0.894	+11.9%	Still too conservative

Model	Pros	Cons
GLM	• Strong transparency and explainability • Well understood validation	• May miss nonlinearities and interaction effects • May require more feature engineering to compete w/ FFNN
FFNN	• Handles nonlinearities and interactions naturally • Embeddings make high-cardinality factors practical	• Governance overhead is higher; explicit calibration checks • Tail behaviour can be weaker unless explicitly addressed
Hybrid A	• Governance-friendlier hybrid • Clean attribution sroy: improvement attributed to severity layer	• Still operationally “two-model” • Can forgo potential serveriy gains from FFNN
Hybrid B	• Keeps incidence highly interpretable • Allows FFNN to add value on severity	• If GLM incidence is miscalibrated, it can dominate PP adequacy • May deliver less aggregate benefit than Hybrid A

Segment adequacy heatmap

Pricing adequacy must be demonstrated at the segment level but ignore one-off noise.



Suggested governance pack

Key objective is to turn model outputs into a repeatable, audit-ready decision.

- **One control sheet for each run**
  - Run ID
  - Dataset / version
  - Owner
  - Reviewer
  - Notes

- **Minimum content**
  - Scope
  - Data & split controls
  - Model specification
  - Validation (test window)
  - Adequacy by segment
  - Stability
  - Operational readiness
  - Attachments

A governance one-pager (control sheet) available in the project GitHub repository (cf. Appendix)

06 Conclusions

Conclusions

For practising actuaries, the most credible path forward may be a hybrid modelling framework in which neural networks are introduced selectively, validated rigourously, and governed as structrured extensions of the actuarial toolkit.

- ✅ GLMs remain strong defaults—they provide the actuarial anchor
- ✅ FFNNs can add value when they pass actuarial validation—they provide targeted predictive enhancement
- ✅ A hybrid framework offers the strongest balance of performance, explainability, and governance

Appendix

References

- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*: MIT Press https://doi.org/10.1007/s10710-017-9314-z
- Holvoet, F., Antonio, K., & Henckaerts, R. (2025). *Neural Networks for Insurance Pricing with Frequency and Severity Data: A Benchmark Study from Data Preprocessing to Technical Tariff*. North American Actuarial Journal, 29(3), 519–562. https://doi.org/10.1080/10920277.2025.2451860
- Richman, R. (2022). *Mind the gap – safely incorporating deep learning models into the actuarial toolkit*. British Actuarial Journal, 27, e21. https://doi.org/10.1017/S1357321722000162
- Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., & Müller, K.-R. (Eds.). (2019). *Explainable AI: Interpreting, explaining and visualizing deep learning*. Springer. https://doi.org/10.1007/978-3-030-28954-6
- Wüthrich, M. V., Richman, R., Avanzi, B., Lindholm, M., Maggi, Marco, Mayer, M., Schelldorfer, J,, and Scognamiglio, S. *AI Tools for Actuaries (January 20, 2026)*. http://dx.doi.org/10.2139/ssrn.5162304

Further research and extensions for practitioners

- Compare Tweedie vs FFNN + tail-aware loss vs two-part models (body + EVT/mixture tail) and measure impact on P90+ MAE
- Evaluate Platt/logistic recalibration, isotonic, and segment-aware credibility calibration (region/ season / occupation) under rolling windows
- Track heatmap persistence over time (raw vs credibility-filtered) and quantify dispersion
- Formalise when to choose Hybrid A vs Hybrid B
- Add local explanations (e.g., SHAP on dense features + embedding diagnostics) with an audit trail
- Stress test against more realistic operational effects (claim reporting lag, inflation regime shifts, benefit cap changes)
- Evaluate whether hazard proxies improve stability/segmentation

GitHub

Visit the GitHub repository to view code and add your own contributions!