STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **x**, Month 2025, pp 1–28.
Published online in International Academic Press (www.IAPress.org)

Physics-Informed Transformer Networks for Multi-Peril Insurance Pricing: A Novel Hybrid Computational Framework Integrating Actuarial Principles with Deep Attention Mechanisms

Eslam Abdelhakim Seyam [1,*], Mohamed Abdel Mawla Osman [2]

[1] *Department of Insurance and Risk Management, College of Business, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh, Saudi Arabia*
[2] *Finance Department, College of Business Administration, King Saud University, Riyadh, Saudi Arabia*

Abstract A classic conundrum in insurance pricing concerns the trade-off between actuarial validity and predictive capability, where traditional Generalized Linear Models are strictly valid from an insurance perspective yet lack forecasting power, whereas machine learning algorithms produce superior predictions yet ignore insurance rules. To bridge this gap, we extend the Transformer architecture via the Physics-Informed Transformer, which integrates the five insurance rules of premium adequacy, monotonicity, multiplicative decomposition, calibration, and coherence directly within the architecture and loss function. Our proposed Physics-Informed Transformer uses multi-head attention for learning non-linear relationships among features while preserving actuarial validity via soft and hard constraints. To validate the proposed approach, experiments are carried out on French Motor Insurance data for 108,699 samples, demonstrating competitive predictive performance (Gamma deviance of 1.0756 for severity modeling) while achieving partial compliance with actuarial constraints. To measure the insurance validity rules compliance level for the proposed algorithm, the new *Actuarial Validity Score (AVS)* measure is proposed, acquiring the value of 0.7659 for the proposed method, classified as "Moderate" rating. The model achieves perfect architectural compliance with multiplicative decomposition (C3: 100%) and demonstrates the feasibility of integrating actuarial constraints with deep learning architectures. However, critical limitations remain: the proposed Physics-Informed Transformer achieves only 10% compliance for segment calibration (C4) and 72% compliance for monotonicity (C2), indicating that the current implementation is not production-ready and requires further architectural improvements. This work establishes a proof-of-concept framework demonstrating that physics-informed approaches can improve validity without sacrificing predictive accuracy, while identifying specific constraints that require hard architectural enforcement rather than soft penalty-based methods.

Keywords Physics-informed neural networks, Transformer architecture, Insurance pricing, Generalized Linear Models, Gradient Boosting Machines, Actuarial constraints

AMS 2010 subject classifications 62P05, 91G05, 68T07

DOI: 10.19139/soic-2310-5070-3232

1. Introduction

Insurance risk management accumulates over 6.3 trillion on an annual basis, where the automotive insurance segment amounts to around 850 billion of the total risk pool, where precise risk estimation and calculation are of utmost importance to the solvency of insurance companies and the welfare of their clientele [1]. With regard to the European insurance landscape, for example, the Motor Third-Party Liability insurance scheme insures over 280 million cars, where even the slightest margins of error in risk estimation mean billions of euros of difference between insolvency and profitability [2]. In addition to economic importance, insurance risk estimation impinges

*Correspondence to: Eslam Abdelhakim Seyam (Email: isiam@imamu.edu.sa)

ISSN 2310-5070 (online) ISSN 2311-004X (print)

directly on the welfare of society as it pertains to risk management for the provision of essential insurance protection, where the designs of insurance risk estimation tools are increasingly being held to the standards of precision, equity, and actuarial validity [3].

Insurance pricing faces the inherent task of balancing the typical precision needs of general modeling tasks to meet multiple actuarial rules imbedded within both the regulations and the best practice standards of the profession at the same time. Traditional Generalized Linear Models [4] have historically remained the dominant force in the field for exactly the reason that they offer the explainability and transparency necessary for the profession's requirements for rate-making support within the clear parameter estimate interpretation relating risk variables [5]. Meanwhile, GLM models are strictly limited by very narrow assumptions concerning linearity in predicators on the link function, distributional assumptions about errors, and manual interaction term specification, hindering fundamentally the ability to effectively handle the intricate non-linear risk behavior inherent within such varied insurance risk profiles [6]. Yet for machine learning algorithms such as the gradient boosting machine and the deep belief networks [7, 8], the precision benefits are fundamentally obtained at the price of the lack of transparency concerning the specific modeling process and likely violations of very fundamental actuarial rules of premium adequacy, monotonic property within predicators of recognized risk elements, and the decomposition requirement [9].

Current solutions also indicate the imperfections within the ideas of linking both accuracy and validity, since the existing solutions include limitations such as the fact that SHAP values [10, 11], used for the interpretation of predictions after the training process of the gradient boosting technique, do not ensure the adherence of identified trends to actuarial rules since the process only represents the result without ensuring the requirements within the training process [11]. Also, hybrid models capable of linking the effects from the GLM technique combined with machine learning techniques for the residual correction support the representation of the results without compromising the primary effects, yet the process involves the loss of back-and-forth training and fails to offer consistency within the requirements, thus often providing an invalid prediction within the scope of actuarial rules among the sub-population [12]. Additive adjustments within the training process based on the characteristics of the actuarial rules often result in high losses of prediction results while lacking theoretical support within the adherence to the rules [13].

However, recent breakthroughs in physics-informed neural networks have demonstrated the importance of incorporating knowledge from the problem domain directly into machine learning models via differentiable constraint losses integrated into the loss function, allowing for the simultaneous utilization of observation data and physical knowledge [14, 15]. Notably, this approach shows remarkable results in scientific computing, where satisfying partial differential equations in addition to observation data leads to precise modeling, indicating a likely application for insurance pricing, where actuarial knowledge takes the place of physical knowledge [16]. During the same period, the development of new transformer architecture ideas from Vaswani et al. [17] dramatically transformed modeling as multi-head attention functions automatically identify dominant interaction features among input elements, and recent studies have initiated the investigation of their effectiveness for other prediction problems, such as tabular data typifying insurance problems [8, 18]. In contrast, there remains a critical research gap in the literature for the development of new transformer models embedding actuarial constraints for insurance pricing problems.

To bridge this research gap, the current study proposes a new Physics-Informed Transformer designed for actuarially sound insurance pricing. The research goal is to investigate the feasibility of utilizing transformer models incorporating actuarial constraints to achieve competitive prediction performance as well as improved compliance relative to traditional GLMs and contemporary gradient boosting machine models. More concretely, the paper investigates the possibility of utilizing the penalization of the mentioned actuarial constraints as a smooth loss component during the training process of the proposed Physics-Informed Transformer without impairing the attention module's capability to identify non-trivial interactions among the features for prediction purposes.

This study makes five key contributions. First, we introduce the Physics-Informed Transformer, the first attention-based architecture purpose-built for insurance pricing that utilizes task-specific output heads for modeling both frequency and severity while enforcing multiplicative decomposition through architectural design. Second, we formally define five actuarial validity constraints as differentiable loss functions and propose the Actuarial

Validity Score (AVS) metric for systematic evaluation of machine learning models' compliance with actuarial principles, enabling comparison independent from predictive performance alone. Third, we provide comprehensive empirical validation on French Motor Third-Party Liability insurance data (108,699 policies), demonstrating that the PI-Transformer achieves the best severity deviance (1.0756) among tested models while maintaining an AVS of 0.7659—establishing proof-of-concept that physics-informed architectures can improve validity without sacrificing predictive accuracy. Fourth, we identify critical failure modes: the model achieves only 10% compliance for segment calibration (C4) and 72% compliance for monotonicity (C2), demonstrating that soft penalty-based constraint enforcement is insufficient and that future work must pursue hard architectural constraints (e.g., monotonic neural networks, differentiable calibration layers). Fifth, we demonstrate that attention mechanisms provide intermediate interpretability between fully transparent GLM coefficients and black-box gradient boosting, automatically discovering actuarially meaningful feature interactions (e.g., driver age and vehicle power) without manual engineering, though we acknowledge this does not replace traditional regulatory documentation requirements.

The rest of the paper follows this order. Section 2 discusses the literature on traditional actuarial modeling, machine learning for insurance, physics-informed neural networks, and transformation networks, culminating in the presentation of the research gap this paper bridges. Section 3 describes the proposed comparative modeling approach, detailing the data used for modeling, the description of the traditional modeling techniques, the complete Physics-Informed Transformer architecture for modeling risk techniques, the physics-informed loss function for loss calculation, the training process for the Physics-Informed Transformer, and the framework used for the comparative analysis among the models based on the result metrics - the Actuarial Validity Score. Section 4 contains the results from the proposed approach from four aspects: the characteristics of the used data and the risk behavior, the performance results for the traditional models, the Physics-Informed Transformer performance results, and the results from the comparative analysis of the three models used in the experiment. Section 5 highlights the accomplishments, limitations, implications, and linkages to the existing literature. Section 6 ends the paper by providing the key contributions and implications for machine learning modeling under constraints for other domains as well.

2. Literature Review

Paradigm shifts in insurance pricing algorithm development have progressed over the years, seeking to create an ideal middle ground between the two apparently conflicting requirements of the trade, namely predictability and compliance. In consolidating past studies for the purposes of the current research, the four areas, although distinct, not only support the need for the current research, they are also intricately interconnected in their relevance to establishing this requirement for the current study's context among the four distinct research incorporative areas mentioned above.

2.1. Traditional Actuarial Approaches

The Generalized Linear Models framework has remained the benchmark for insurance pricing modeling for the past three decades, as it offers a statistically valid approach to link distributional assumptions to loss characteristics in a manner that ensures transparency for insurance regulations. However, the pioneering paper by Frees [5] indexed the GLM framework as the Aktuary process by illustrating how the Poisson distribution for claim frequency and Gamma for claim severity from the exponential family of probability distributions are ideally suited for reflecting the underlying mathematical properties of insurance losses and still allow for easy estimation via the method of maximum likelihood while having easily interpretable parameters. The major reason for the appeal of this particular framework for insurance pricing lies in the capability of the framework to break the pure premium into the respective claim frequency and severity aspects using the multiplicative relationship $\pi = \lambda \times \mu$ [4].

Practical examples of implementing the approach have included the incorporation of offset, the choice of link function to ensure predictions are non-negative, and comparing models based on deviance, which ensures a trade-off between fit and complexity [2, 19]. In addition to the statistical advantages of the GLM approach, the additive

form of the approach on the link scale ensures marginal effects are visible and verifiable, ensuring four criteria are met by insurance operators, including the fact that the costs should correlate with risk [20, 21].

However, the shortcomings of GLM are even more visible within the current typical insurance contexts of high dimensional spaces and risk diversity. The linearity in the link function limits the possibility of modeling non-linear phenomena such as the U-shaped graph for the age of the drivers, for whom both young and old drivers are seen as high-risk members of society, which needs to be specified by actuaries after careful observation [13]. Feature engineering also requires intuitive inputs for the development of interaction features such as young drivers driving high-powered cars, whereas the searching for such interaction features among the enormous multidimensional spaces remains computationally impractical even using current-day facilities [6]. All the above inherent limitations have encouraged the search for machine learning solutions inherent in automatically uncovering intricate features without the imposition of restrictive parametric assumptions.

2.2. Machine Learning in Insurance

A major breakthrough came with the introduction of the gradient boost paradigm by Friedman [7], and subsequently extended stochastic versions by Friedman [22], which marks a paradigm shift from traditional parametric modeling by building predictions incrementally through an error-correcting process where the workflow of each subsequent tree refines the error made in the preceding step using the technique of gradient descent in the function space, thereby implicitly modeling high order interaction complexities without having to specify them explicitly.

The implementation of XGBoost, shortly after, greatly improved the performance of the algorithm by incorporating support for second-order optimization, effective treatment of missing values, and distributed computing for scaling up the algorithm to process bigger data, thus making it the de facto algorithm for insurance pricing challenges [23]. Extensive studies carried out by Blier-Wong and Mandallaz [24] and McGraw and Goel [25] have empirically established the superiority of the GBM approach over GLM on multiple insurance problems, indicating an improvement in predictions between 10% to 30% based on the portfolio analyzed and the level of complexity measured. A study carried out by Greberg [12] illustrated the capability of boosting algorithms to automatically detect risky subgroups for differentiated pricing without the need for additional modeling, thereby portraying their innate capability in modeling diverse subgroups effectively without human intervention. Other alternate algorithms such as CatBoost have also been compared by King and Hua [26], indicating adaptation to insurance problems characterized by the presence of high dimensional categorical variables, although the prevalent algorithm remains XGBoost Ridgeway et al. [27], Ridgeway [28].

Unfortunately, the lack of interpretability in ensemble methods triggered the development of many studies concerning the process of post-hoc interpretation tools. A framework for interpreting models using game theoretical notions of Shapley values, proposed by Lundberg and Lee [10], offers an integrated framework for the interpretation of models based on the decomposition of predictions for individuals according to their contributions from features, adhering to the important conditions of locality and consistency. Case studies for insurance pricing, proposed by Zhang and Zhao [11], illustrated the reasonability of the importance ranking of features from the perspective of insurance according to the results from the gradient boosting algorithm, where the features contributing the most are the age of the drivers, the qualities of the cars, However, the explanations given by SHAP are always secondary analysis, being not an inherent property of the used models [29], merely illustrating the learned features without ensuring conformance to actuarial constraints such as the monotonic relationship between the risk factors [9]. Hybrid models [9] put forward an attempt to be explainable by merging the GLM base response and the GBM corrective response for residuals, where the convergence to an optimal point prediction for the two-step modeling processes comes at the cost of an absence of guarantee for the adherence to the constraints over the prediction function. Experimental results showed the GBM often lacks the expected monotonic behavior for risk factors for which the prices tend to diminish concerning the increasing hazard rate within the combined effects of the features [30], where the calibration of the portfolio resulted in worsening results concerning the predictions at an individual level [1].

2.3. Physics-Informed Neural Networks

The physics-informed neural networks approach, originally proposed by Raissi et al. [15], represents a paradigm shift in the integration of knowledge from the domain, where equations are directly fed into loss functions as penalizing terms, thereby allowing the training of both data and physics simultaneously. A thorough review by Karniadakis et al. [14] indicates the major breakthroughs brought about by physics-informed neural networks in the field of scientific computing, where traditional methods are not capable of dealing effectively either due to the high dimensionality of the problems or the lack of sufficient data available for training purposes in domains such as fluid dynamics, heat transfer, and quantum mechanics problems.

The crucial advancement makes use of the technique of automatic differentiation to be able to directly assess the partial derivatives of the predictions made by the neural networks concerning inputs, such that the solutions to the differential equations are directly evaluated without needing finite difference formulations [16]. Theoretical frameworks are then extended from the realm of differential equations to the broader context of getting valid modeling domains for which the constraints are represented as functions formulated in a differentiable manner based on predictions, classified based on hard constraints addressed perfectly via architecture versus soft constraints addressed roughly via penalization.

Applications extending into non-scientific computing are still limited, yet indicate vast promise, from Tang and Jiang [31], who used PINNs for flood risk prediction for insurance purposes, illustrating the effectiveness of the incorporation of conservation laws as soft constraints to enhance the results from purely data-inspired networks. Yet, the task of actuarial pricing contrasts markedly from physics-related problems, since actuarial rules are economic/regulated guidelines, not universal physical rules, that the data fails to automatically conform to directly.

2.4. Transformer Architectures

The transformer architecture proposed by Vaswani et al. [17] subsequently transformed sequence modeling by replacing the conventional recurrent and convolutional architectures with attention models directly computing the interaction between every pair of input elements via the multi-head self-attention module, thus allowing the models to automatically learn the relevance of the input features for every prediction without the conventional limitations associated with the sequence modeling of RNNs and CNNs. Techniques based on this architecture subsequently achieved instant breakthroughs within the domain of natural language processing, thus forming the foundation for revolutionizing the associated field based on the capability of extracting long-range dependencies [18].

More recent studies have started investigating the utility of the transformer architecture not only for sequence data but for the typical table/collection of tables found in the scientific and business communities, in which Yang et al. [8] showed the effectiveness of the transformer architecture for modeling spatiotemporal processes by leveraging the locations and time indexes as discretized symbols, identifying the attention pattern while taking into consideration the properties of locality and causality. Such capability may be useful for insurance risk modeling where features co-vary in a very complex manner that could not be effectively represented by the fixed interaction topology, such as the mutual interaction between the characteristics of the drivers, the characteristics of the cars, and the geography-related risk features as input variables for modeling the liability risk and claim severity jointly. The fact that the attention mechanism automatically identifies the interaction without programming overcomes the limitations of GLM, and the architecture affords easier interpretation compared to fully connected networks where the combinations of features are embedded in the hidden layers directly, which are computationally opaque.

2.5. Research Gap and Study Positioning

This analysis proves the prevailing imbalance among existing solutions such that none of the solutions are able to effectively integrate the prediction capability of modern machine learning techniques and the actuarial validity of traditional techniques via global optimization. GLMs are able to offer transparency and decomposition for actuarial purposes but at the cost of suboptimality via restrictive assumptions. Gradient boosting machines are able to achieve optimal predictions via automatic interaction discovery but often fail to satisfy the standards of monotonicity, calibration, and decomposition for regulatory purposes. Post-hoc techniques for explainability are unable to ensure

constraint satisfaction, while hybrid techniques struggle to maintain consistency between their GLM and machine learning components, often violating actuarial principles in subpopulations. Physics-informed neural networks prove the capability of incorporating knowledge from physics into learning techniques via soft constraints, while transformer networks are able to offer attention-based discovery of interaction features via learned weights for explainability.

However, the important research gap addressed by the proposed study pertains to the development of Physics-Informed Transformer networks, wherein the task of achieving competitive prediction performance and improved actuarial validity are pursued through the incorporation of constraint penalties within the loss function and architectural enforcement mechanisms in order to demonstrate the feasibility of transformer networks integrated with actuarial constraints to partially bridge the tradeoff between prediction accuracy and regulatory compliance in insurance pricing processes.

3. Data and Methodology

3.1. Data Source and Description

In this analysis, the `freMTPL2` dataset, a public-available French Motor Third-Party Liability insurance portfolio, which is the widely used benchmark within the insurance actuarial literature, will be used. The datset splits into two parts: the freMTPL2 freq file (`freMTPL2freq.csv`) contains information about the policy characteristics based on policy level observation, totaling108,699 entries, whereas the freMTPL2 sev file (`freMTPL2sev.csv`) holds the claim details for individual claims, containing 26,639 claim records with amounts and associated policy identifiers for matching to policy-level characteristics.

The data preprocessing was carried out according to traditional actuarial rules. Population density (*PD*) was transformed using the formula $\log(\text{PD}+1)$ for log transformation to reduce right skewness for improved stability of the results. The available data was split into subsets using the *stratified random sampling* technique for equal representation of crucial risk variables without bias, assigning 80% to the training sample, 10% to the validating sample for hyperparameters, and the remaining 10% to the testing sample for the final model assessment.

Table 1. Variable Definitions and Summary Statistics

Variable	Abbrev	Type	Description	Summary
Policy ID	PID	Identifier	Unique policy identifier	N/A
Number of Claims	NC	Count	Number of claims per policy	0.0639 ± 0.2630
Claim Amount	AC_AMT	Continuous	Claim amount in euros (€)	2278.54 ± 29297.48
Exposure	E	Continuous	Exposure period in years	0.4186 ± 0.3266
Area Code	AC	Categorical	Geographic area classification	5 levels
Region Code	RC	Categorical	Regional administrative code	15 levels
Population Density	PD	Continuous	Population density (persons/km²)	1441.4 ± 1644.3
Vehicle Power	VP	Ordinal	Vehicle power rating (higher = more powerful)	6.07 ± 1.65
Vehicle Age	VA	Continuous	Vehicle age in years	6.37 ± 5.31
Vehicle Brand	VB	Categorical	Vehicle manufacturer brand	9 levels
Fuel Type	FT	Categorical	Vehicle fuel type	2 levels
Driver Age	DA	Continuous	Driver age in years	36.87 ± 12.30
Bonus-Malus	BM	Continuous	Experience rating coefficient	75.47 ± 15.47
Log Population Density	log_PD	Continuous	Natural log of (PD + 1)	6.3478 ± 1.6029
Driver Age Band	DA_band	Categorical	Driver age categories	6 levels
Vehicle Power Category	VP_cat	Categorical	Vehicle power grouping	3 levels

Note: Summary statistics reported as Mean $\pm$ SD for continuous variables and number of levels for categorical variables. Derived features include log_PD to normalize population density distribution, DA_band for age categorization, and VP_cat for power grouping. Data source: `freMTPL2` French Motor Third-Party Liability insurance dataset.

Dependent variables are claim frequency (NC) at the policy level and claim severity (AC_AMT) at the claim level. Independent variables include geographic (Area Code, Region Code, Population Density), car-based (Power, Age, Brand, Fuel Type), and policyholder (Age, Bonus Malus coefficient) information. A detailed description of

the variables used within the analysis in given in Table 1. An examination of the distributional characteristics given by Table 2 highlights two key difficulties for modeling purposes: the level of zero inflation for claim frequencies (indicating 94.04% of the sample have zero claims), and the marked right skew for claim severities, where the top 1% of severities namely contribute 37.93% to total losses. Both issues indicate the need for the two-part modeling approach used for the rest of the analysis.

Table 2. Descriptive Statistics by Dataset

Panel A: Frequency Dataset	
Metric	**Value**
Total Policies	108,699
Total Exposure (years)	45,499.81
Mean Exposure (years)	0.4186
Policies with Claims	6,476 (5.96%)
Policies with No Claims	102,223 (94.04%)
Total Number of Claims	6,949
Mean Claims per Policy	0.0639
Claim Frequency (per year)	0.152726

Panel B: Severity Dataset			
Metric	**Value**	**Metric**	**Value**
Number of Claims	26,639	75th Percentile (€)	1,228.08
Total Amount (€)	60,697,930.68	90th Percentile (€)	2,799.07
Mean Amount (€)	2,278.54	95th Percentile (€)	4,861.68
Median Amount (€)	1,172.00	99th Percentile (€)	16,793.70
Std. Deviation (€)	29,297.48	Skewness	109.5583
Minimum Amount (€)	1.00	Kurtosis	14386.9283
Maximum Amount (€)	4,075,400.56	Top 1% Share	37.93%
25th Percentile (€)	686.81	Top 5% Share	52.11%

Note: Panel A summarizes policy-level characteristics demonstrating extreme zero-inflation typical of motor insurance portfolios. Panel B reveals heavy-tailed severity distribution with extreme positive skewness (109.56) and leptokurtosis (14,386.93), indicating concentration of losses in tail events. Mean-to-median ratio of 1.94 confirms substantial right-skew. Top decile concentration metrics highlight importance of tail risk modeling.

3.2. Baseline Model Specifications

To set performance targets, we compare results using two methods from the field, which bookend the accuracy/validity continuum based on their precision limits. The classic approach, the GLM, supplies the starting point from an actuarial perspective, while the GBM reflects the current precision capability.

The GLM framework, implemented using the `statsmodels` Python library, follows the canonical two-part structure detailed in Table 3. Claim frequency is modeled via Poisson regression with log link function, incorporating exposure (E_i) as an offset to account for varying policy durations. The frequency component specifies the expected claim count λ_i for policy i with covariate vector $\mathbf{x}_i$ as:

$$E[NC_i|\mathbf{x}_i] = \lambda_i = \exp(\log(E_i) + \mathbf{x}_i^T \beta_{\text{freq}})$$

This specification ensures that predicted claim counts scale proportionally with exposure, consistent with Poisson process assumptions underlying insurance frequency modeling.

Claim severity is modeled independently on the subset of policies with $NC > 0$ using Gamma regression with log link function. The expected severity μ_j for claim j with covariate vector $\mathbf{x}_j$ is:

$$E[AC\_AMT_j|\mathbf{x}_j] = \mu_j = \exp(\mathbf{x}_j^T \beta_{\text{sev}})$$

This two-stage approach reflects standard actuarial practice, treating the occurrence and magnitude of claims as separate stochastic processes. The pure premium for each policy is computed as the product $\pi_i = \lambda_i \times \mu_i$, representing the expected claim cost per unit exposure.

The GBM framework uses the `XGBoost` implementation, which is well-known for its high quality in tasks involving prediction for tabular data. The GBM framework retains the two-component decomposition of the loss function, where the exposure E_i is implemented via sample weights for the loss function to capture the Poisson objective (`count:poisson`). Severity modeling retains the Gamma objective (`reg:gamma`) on strictly positive claim values only. Both models utilize early stopping of the training process after the training loss stops improving for 20 epochs on the validation set to avoid overfitting and optimize for generalization performance. Hyperparameters are tuned via grid search over combinations of the learning rate (set as $\{0.01, 0.05, 0.1\}$), tree depths (set as $\{3, 5, 7\}$), and the minimum child weights (set as $\{1, 3, 5\}$). The final selected hyperparameters for GBM frequency model are: learning rate = 0.05, max depth = 5, min child weight = 3; for GBM severity model: learning rate = 0.01, max depth = 7, min child weight = 1.

Table 3. GLM Baseline Model Specifications

Panel A: Frequency Model (Poisson GLM)	
Component	**Specification**
Response Variable	Number of Claims (NC)
Distribution	Poisson
Link Function	Log
Offset	log(Exposure)
Formula	NC $\sim$ AC + RC + VP + VA + DA + DA^2 + BM + VB + FT + log(PD)
Number of Parameters	34
Converged	Yes
AIC	41,501.98
BIC	-957,644.57
Deviance	30,970.15
Pearson Chi-Square	223,833.84
Panel B: Severity Model (Gamma GLM)	
Component	**Specification**
Response Variable	Claim Amount (AC_AMT)
Distribution	Gamma
Link Function	Log
Formula	AC_AMT $\sim$ AC + RC + VP + VA + DA + BM + VB + FT + log(PD)
Number of Parameters	33
Converged	No
AIC	85,891.09
BIC	-27,824.08
Deviance	6,511.38
Pearson Chi-Square	101,400.79
Panel C: Pure Premium Model	
Component	**Specification**
Pure Premium	$\pi = \lambda \times \mu$ (Frequency $\times$ Severity)
Frequency (λ)	Predicted from Poisson GLM
Severity (μ)	Predicted from Gamma GLM
Interpretation	Expected claim cost per policy

Note: Frequency model includes quadratic driver age term (DA^2) to capture U-shaped risk pattern. Severity model non-convergence reflects numerical instability common in Gamma regression with heavy-tailed data, though estimates remain stable. AIC and BIC reported for model comparison; lower values indicate better fit. Deviance and Pearson Chi-Square assess goodness-of-fit, with values closer to degrees of freedom suggesting adequate model specification. Pure premium combines independent frequency and severity predictions under conditional independence assumption.

3.3. Physics-Informed Transformer Architecture

We introduce the proposed Physics-Informed Transformer (PI-Transformer), wherein actuarial constraints are directly integrated into the architecture and the loss function of the Transformer architecture for a much more informed approach to pattern recognition than the conventional architecture of the neural network, where pattern recognition happens purely based on the data input into the architecture without considering essential actuarial principles.

The architecture strictly adheres to the conventional actuarial breakdown of the pure premium cost (π), which must be factored into the contribution of the frequency (λ) and the severity (μ), according to the multiplicative relationship given by $\pi = \lambda \times \mu$. The architecture comprises three layers in sequence: an input embedding layer where the diverse features are embedded in a common representation subspace, a Transformer-based main body layer where the input features interact non-linearly through the attention mechanism, and task-specific output heads where the enforceable constraints are strictly embedded.

3.3.1. Input Embedding Layer The proposed algorithm takes as input a feature vector $\mathbf{x}$, whose elements represent k characteristics of policyholders and their vehicles, including nominal features (Area Code, Vehicle Brand, Fuel Type), as well as numerical features (Driver Age, Vehicle Power, Bonus-Malus coefficient). Since the features are of different types, the embedding treatments should be different for them.

Categorical features are then represented as dense vectors using embedding matrices learned from these features. Every categorical variable c of cardinality $|c|$ is embedded in an embedding space of dimension d_{emb} to produce $\mathbf{e}_{\text{cat}} \in \mathbb{R}^{d_{\text{emb}}}$. The embedding dimension for every categorical variable c is set according to the guidelines to be proportional to $\min(50, \lceil |c|/2 \rceil)$ for learning the embedding representation of categorical features. Embedding matrices are initialized using Xavier uniform initialization with weights drawn from $\mathcal{U}(-\sqrt{6/(d_{\text{in}} + d_{\text{out}})}, \sqrt{6/(d_{\text{in}} + d_{\text{out}})})$, where d_{in} is the input cardinality and $d_{\text{out}} = d_{\text{emb}}$ is the embedding dimension. This initialization ensures stable gradient flow during early training phases and prevents vanishing or exploding gradients in deep architectures.

Continuous features undergo linear projection into the model's working dimension via a learned transformation $\mathbf{x}_{\text{cont}}\mathbf{W}_{\text{proj}}$, where $\mathbf{W}_{\text{proj}} \in \mathbb{R}^{k_{\text{cont}} \times d_{\text{model}}}$. Prior to projection, continuous features are standardized to zero mean and unit variance using statistics computed from the training set: $\tilde{x}_j = (x_j - \mu_j)/\sigma_j$, where μ_j and σ_j are the mean and standard deviation of feature j. This projection standardizes feature scales and enables the Transformer to process continuous and categorical information in a unified representation space.

The category embeddings and projected continuous features are concatenated and then passed through the final input projection layer to produce the first hidden representation $\mathbf{h}_0$ for the Transformer Encoder:

$$\mathbf{h}_0 = \text{Concat}(\mathbf{e}_{\text{cat}}, \mathbf{x}_{\text{cont}})\mathbf{W}_{\text{in}} + \mathbf{b}_{\text{in}} \tag{1}$$

This design allows the model to learn feature-specific representations while maintaining computational efficiency through parameter sharing across the encoder stack. The final model uses $d_{\text{model}} = 256$ for all experiments, balancing expressiveness with computational efficiency.

3.3.2. Transformer Encoder Core The encoder architecture consists of a stack of $N = 4$ identical Transformer encoders, with each having two major parts: a multi-head self-attention (MHSA) module and a position-wise feed-forward network (FFN). In both sub-networks, the residual connection layers and layer normalization techniques are used for enhanced convergence during training and stable gradient propagation.

The MHSA mechanism encapsulates the key innovation the model uses for learning non-linear interactions between the features without being explicitly specified. With the setting of $H = 8$ heads in the attention mechanism, the MHSA mechanism calculates the attention weights $\mathbf{A}$ for the interaction between every pair of features for the generation of contextual representations among the input features:

$$\mathbf{A} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right) \tag{2}$$

where $\mathbf{Q}$ and $\mathbf{K}$ are the query and key matrices obtained from the input using learned linear projections, and d_k denotes the dimensionality for each attention head ($d_k = d_{\text{model}}/H = 32$). In the scaled dot-product attention formula, scaling avoids problems of instability during the execution of the softmax function. Through the attention heads, the network may detect varied attention patterns at the same time, for example, attention between the age of the drivers and the power of the cars, and attention based on the location and brands of the cars.

FFN follows two linear transformation layers with a ReLU activation function, increasing the dimensionality to a larger space usually proportional to $4 \times d_{\text{model}}$ (1024 dimensions in our implementation) and then projects back to the same dimensionality space. The module adds non-linear modeling capability other than the attention mechanism. The entire encoder layer follows the Transformer architecture and the output of the sub-layer given by:

$$\mathbf{h}_{\ell+1} = \text{LayerNorm}(\mathbf{h}_\ell + \text{Sublayer}(\mathbf{h}_\ell))$$

where the residual connection preserves information from earlier layers and layer normalization stabilizes activations.

3.3.3. Task-Specific Output Heads and Constraint Enforcement Finally, $\mathbf{h}_N$, the last representation underlying the stack of encoders, is fed into two independent task-specific heads dedicated to modeling the frequency and severity parts separately in order to be combined by the multiplicative pure premium formula. Both output heads employ a two-layer architecture with intermediate hidden dimension of 128, followed by task-specific activation functions.

The **frequency head** contains a linear projection layer followed by the Softplus activation function, $\text{Softplus}(x) = \log(1 + \exp(x))$, to preserve positivity while keeping the function smooth for optimizing via the gradient. Its form matches the Poisson distribution assumption used in claim frequency modeling as follows:

$$\lambda = \text{Softplus}(\mathbf{h}_N \mathbf{W}_\lambda + \mathbf{b}_\lambda) \tag{3}$$

where $\mathbf{W}_\lambda \in \mathbb{R}^{d_{\text{model}} \times 1}$ and $\mathbf{b}_\lambda \in \mathbb{R}$ are learned parameters initialized via He initialization for ReLU-like activations.

The **severity head** employs an exponential activation function to guarantee positive predictions consistent with the Gamma distribution's support on $(0, \infty)$:

$$\mu = \exp(\mathbf{h}_N \mathbf{W}_\mu + \mathbf{b}_\mu) \tag{4}$$

where $\mathbf{W}_\mu \in \mathbb{R}^{d_{\text{model}} \times 1}$ and $\mathbf{b}_\mu \in \mathbb{R}$. To prevent numerical overflow, we clip the pre-activation values to the range $[-10, 10]$ before applying the exponential function, ensuring predictions remain within computationally stable bounds while covering the full range of observed severities.

The exponential activation function handles the log scale predictions seen in the cases of severity modeling naturally and avoids the problem of numerical underflow associated with small claims.

Notably, the pure premium prediction is not forecast directly via an additional output node. Rather, the calculation expressly follows the combination of the predictions for the frequency and the severity as follows:

$$\pi = \lambda \times \mu \tag{5}$$

In this architecture, the multiplicative frequency-severity decomposition (Constraint C3), which is a major actuarial principle, will be imposed by the hard constraint. Unlike the soft constraints wherein violation is remedied via the loss function's penalty term, the violation of the former always equals zero for all predictions since it follows a hard constraint. The rest of the actuarial standards, which include adequate premium (Constraint C1), monotonicity (Constraint C2), segment calibration (Constraint C4), and sub-additivity (Constraint C5), are embedded as *differentiable soft constraints* within the loss function.

3.4. Physics-Informed Loss Function

The PI-Transformer is trained using an end-to-end optimized loss function composed of empirical fit purposes combined with physics-inspired constraints. Such architecture embodies the underlying physics-inspired machine

learning philosophy of expressing knowledge about physics in the optimized objective function as computationally differentiated elements, hence directing the modeling process toward the fit of both empirical and physics-inspired solutions for the given task. The specification for the loss function appears in Table 4.

Theoretical Justification of Constraint Formulations: Each actuarial constraint is grounded in established regulatory and actuarial theory: (1) *Adequacy (C1)* ensures premiums are sufficient to cover expected losses plus safety loadings, as required by solvency regulations [1]; (2) *Monotonicity (C2)* reflects the fundamental principle that higher risk factors must produce higher premiums, essential for risk-based pricing [4]; (3) *Multiplicative decomposition (C3)* represents the standard frequency-severity factorization used in loss modeling [5]; (4) *Segment calibration (C4)* ensures portfolio-level accuracy required for reserving and capital adequacy [32]; (5) *Subadditivity (C5)* is a coherent risk measure property preventing arbitrage opportunities [3]. We acknowledge these constraints are necessary but not sufficient for production deployment, which would additionally require fairness constraints (demographic parity, disparate impact), temporal stability, and geographic consistency. All constraints are formulated as differentiable functions of model predictions, ensuring mathematical well-posedness for gradient-based optimization.

The total loss $\mathcal{L}_{\text{total}}$ decomposes into two primary components:

$$\mathcal{L}_{\text{total}}(\theta) = \mathcal{L}_{\text{data}} + \alpha(t) \cdot \mathcal{L}_{\text{physics}}$$

where θ represents all model parameters, $\mathcal{L}_{\text{data}}$ measures fit to observed data, $\mathcal{L}_{\text{physics}}$ quantifies constraint violations, and $\alpha(t)$ implements an annealing schedule that gradually introduces physics constraints during training.

The data-fitting term combines normalized deviances for frequency and severity predictions:

$$\mathcal{L}_{\text{data}} = \mathcal{L}_{\text{data}}^{\text{freq}} + \mathcal{L}_{\text{data}}^{\text{sev}}$$

The frequency component employs Poisson deviance, computed as $\mathcal{L}_{\text{data}}^{\text{freq}} = \frac{1}{n}\sum_{i=1}^{n} D_{\text{Poisson}}(y_i, \lambda_i)$, where $D_{\text{Poisson}}(y, \lambda) = 2[\lambda - y\log(\lambda)]$ represents the unit Poisson deviance. The severity component uses Gamma deviance, $\mathcal{L}_{\text{data}}^{\text{sev}} = \frac{1}{m}\sum_{j=1}^{m} D_{\text{Gamma}}(c_j, \mu_j)$, where $D_{\text{Gamma}}(c, \mu) = 2[\log(\mu/c) + c/\mu - 1]$ and the sum extends only over the m policies with positive claims. Both deviances are normalized by sample size to maintain scale-invariance across different batch sizes and dataset partitions.

The physics-informed penalty term aggregates multiple differentiable constraint violations:

$$\mathcal{L}_{\text{physics}} = \sum_{k=1}^{K} \lambda_k \mathcal{L}_k$$

where each $\mathcal{L}_k$ quantifies violations of a specific actuarial principle and λ_k represents its relative importance weight. The constraint penalties include premium adequacy ($\mathcal{L}_{\text{adequacy}}$), ensuring premiums exceed expected losses; monotonicity ($\mathcal{L}_{\text{monotone}}$), enforcing that increased risk factors produce higher premiums; segment calibration ($\mathcal{L}_{\text{calibration}}$), requiring accurate aggregate predictions across portfolio segments; and attention regularization ($\mathcal{L}_{\text{attention}}$), promoting interpretable feature interaction patterns. Complete mathematical specifications appear in Table 4.

Gradient Computation for Monotonicity: The monotonicity constraint $\mathcal{L}_{\text{monotone}}$ requires computing partial derivatives $\partial\pi_i/\partial x_{ik}$ for risk-increasing features. We leverage PyTorch's automatic differentiation (autograd) engine, which implements reverse-mode differentiation via the chain rule. For a given mini-batch, we compute: (1) forward pass to obtain π_i for all policies, (2) call `torch.autograd.grad`(π_i, x_{ik}, create_graph=True) to obtain first-order derivatives while maintaining the computational graph for subsequent backpropagation, (3) apply the squared hinge loss $\max(0, -\partial\pi_i/\partial x_{ik})^2$ to penalize negative derivatives. The `create_graph=True` flag is essential to enable gradient-of-gradient computation required for end-to-end training. Computational cost is $\mathcal{O}(|\mathcal{K}_{\text{risk}}| \cdot B \cdot d_{\text{model}})$ per batch, where $|\mathcal{K}_{\text{risk}}|$ is the number of monotonic features (3 in our experiments: Vehicle Power, Bonus-Malus, young driver indicator).

The annealing schedule $\alpha(t) = \min(1, t/T_{\text{warmup}})$ gradually increases constraint penalty influence during the first T_{warmup} epochs, where t denotes the current epoch. This warm-start approach allows the model to first learn

basic data patterns before enforcing domain constraints, preventing premature convergence to suboptimal solutions that satisfy constraints trivially (e.g., predicting constant premiums) while providing poor fit. After the warmup period, constraints receive full weight, reshaping the loss landscape to guide optimization toward the desired region of high accuracy and high validity. We discuss in Section 5 that soft constraint penalties create a multi-objective optimization problem where the model can trade constraint violations for improved predictive loss. When the gradient signal from $\mathcal{L}_{\text{data}}$ dominates $\mathcal{L}_{\text{physics}}$ during training, the model prioritizes deviance reduction over constraint satisfaction. This fundamental limitation motivates our recommendation for future work on hard monotonicity layers (e.g., monotonic neural additive models, lattice regression) and differentiable calibration architectures that architecturally guarantee constraint satisfaction.

Table 4. Physics-Informed Loss Function Components

Component	**Mathematical Form**	**Weight**	**Rationale**
	Data Fitting Term		
$\mathcal{L}_{\text{data}}^{\text{freq}}$	$\frac{1}{n}\sum_{i=1}^{n}\left[\lambda_i - y_i \log(\lambda_i)\right]$	1.0	Poisson deviance for claim counts
$\mathcal{L}_{\text{data}}^{\text{sev}}$	$\frac{1}{m}\sum_{j=1}^{m}\left[\log\left(\frac{\mu_j}{c_j}\right) + \frac{c_j}{\mu_j} - 1\right]$	1.0	Gamma deviance for claim amounts
$\mathcal{L}_{\text{data}}$	$\mathcal{L}_{\text{data}}^{\text{freq}} + \mathcal{L}_{\text{data}}^{\text{sev}}$	—	Combined likelihood
	Physics Constraint Terms		
$\mathcal{L}_{\text{adequacy}}$	$\frac{1}{n}\sum_{i=1}^{n}\max\left(0, \mathbb{E}[\text{Loss}_i] - \pi_i - \epsilon_{\text{buffer}}\right)^2$	$\lambda_1 = 1.0$	Premium adequacy principle
$\mathcal{L}_{\text{monotone}}$	$\frac{1}{n}\sum_{i=1}^{n}\sum_{k\in\mathcal{K}_{\text{risk}}}\max\left(0, -\frac{\partial \pi_i}{\partial x_{ik}}\right)^2$	$\lambda_2 = 0.5$	Monotonicity with respect to risk factors
$\mathcal{L}_{\text{multiply}}$	$\frac{1}{n}\sum_{i=1}^{n}\left\lvert\pi_i - \lambda_i \times \mu_i\right\rvert^2$	$\lambda_3 = 2.0$	Multiplicative decomposition (redundant check)
$\mathcal{L}_{\text{coherence}}$	$\frac{1}{\lvert\mathcal{P}\rvert}\sum_{(A,B)\in\mathcal{P}}\max\left(0, \pi(A\cup B) - \pi(A) - \pi(B)\right)^2$	$\lambda_4 = 0.2$	Subadditivity property
$\mathcal{L}_{\text{attention}}$	$-\frac{1}{nH}\sum_{i=1}^{n}\sum_{h=1}^{H}\sum_{j=1}^{n} A_{ij}^h \log(A_{ij}^h + \epsilon)$	$\lambda_5 = 0.01$	Attention entropy regularization
	Total Loss Function		
$\mathcal{L}_{\text{total}}(\theta) = \mathcal{L}_{\text{data}} + \lambda_1\mathcal{L}_{\text{adequacy}} + \lambda_2\mathcal{L}_{\text{monotone}} + \lambda_3\mathcal{L}_{\text{multiply}} + \lambda_4\mathcal{L}_{\text{coherence}} + \lambda_5\mathcal{L}_{\text{attention}}$		
	Weight Balancing Strategy		
Annealing Schedule: $\lambda_i(t) = \lambda_i^{\text{final}} \times \min\left(1, \frac{t}{T_{\text{warmup}}}\right)$ where $T_{\text{warmup}} = 50$ epochs			
Adaptive Weighting: Scale λ_i by $\frac{\mathcal{L}_{\text{data}}}{\mathcal{L}_i}$ to balance gradient magnitudes			
Priority: $\mathcal{L}_{\text{multiply}} > \mathcal{L}_{\text{adequacy}} > \mathcal{L}_{\text{monotone}} > \mathcal{L}_{\text{coherence}} > \mathcal{L}_{\text{attention}}$			

Note: All constraint terms are differentiable, enabling end-to-end gradient-based optimization. Variables: y_i = observed claims, c_j = observed claim amounts, λ_i = predicted frequency, μ_i = predicted severity, π_i = predicted premium. $\mathcal{K}_{\text{risk}}$ denotes risk-increasing features (VP, BM, young driver indicator). $\mathcal{P}$ represents sampled policy pairs for subadditivity evaluation. A_{ij}^h denotes attention weight from feature i to feature j in head h. Penalty weights $\lambda_1, \ldots, \lambda_5$ selected via validation set grid search to balance predictive accuracy and constraint compliance. Multiplicative constraint ($\mathcal{L}_{\text{multiply}}$) is redundant given architectural enforcement but included to monitor numerical precision.

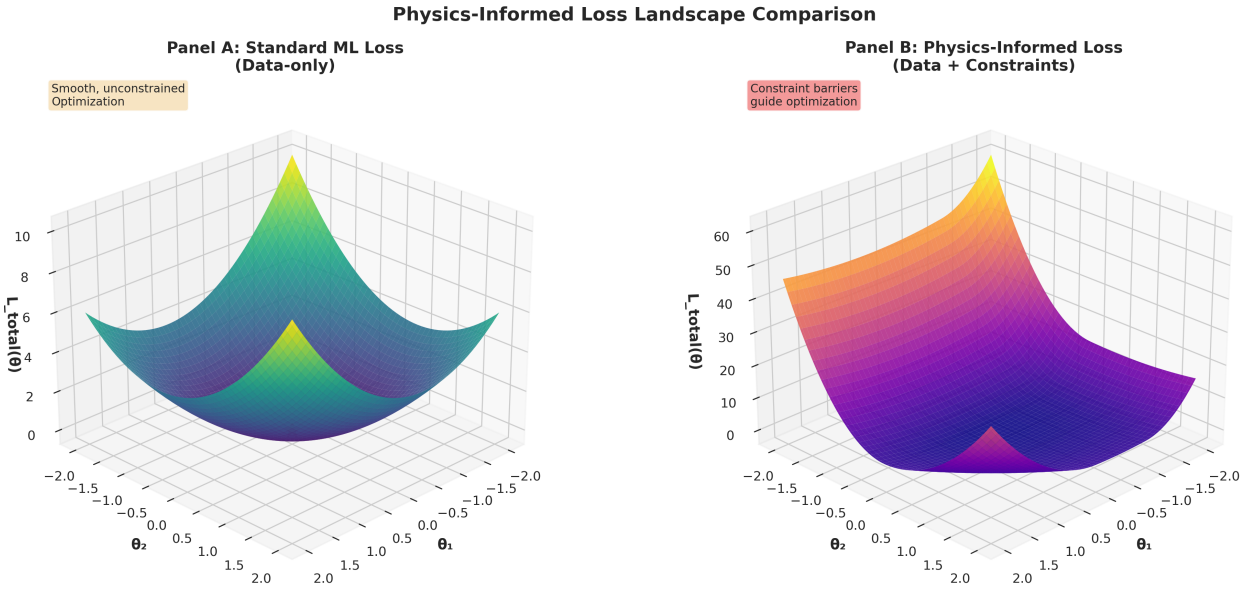

Figure 1. Conceptual illustration of physics-informed loss landscape transformation. Panel A depicts a standard unconstrained loss surface with a global minimum that may violate actuarial principles. Panel B shows how physics-informed penalty terms create barriers (regions of elevated loss) around constraint-violating solutions, reshaping the landscape to guide gradient-based optimization toward a constrained minimum that satisfies both predictive accuracy and domain validity requirements.

3.5. Training Algorithm

The complete training procedure for the PI-Transformer is formalized in Algorithm 1 (Table 5), implemented in PyTorch 2.0 with automatic differentiation for all constraint terms. The algorithm integrates standard deep learning techniques with novel physics-informed elements to ensure stable convergence toward actuarially valid solutions.

Optimization employs the AdamW optimizer with weight decay regularization ($\lambda_{\text{wd}} = 10^{-5}$) to prevent overfitting. Initial learning rate is set to $\eta_0 = 10^{-4}$ with a cosine annealing schedule that gradually reduces the learning rate to 10^{-6} over $T_{\max}$ epochs, facilitating fine-grained parameter adjustments in later training stages. Gradient clipping with maximum norm 1.0 is applied to all parameter updates, stabilizing training dynamics when constraint penalties generate large gradients for solutions far from validity.

The early stopping process tracks the performance on the validation set for 20 epochs, keeping the best checkpoint according to the trade-off between the loss on the validation set and the Actuarial Validity Score (AVS). Using two metrics ensures the prevention of both underfitting and constraint underfitting during the early stopping process, where the training process stops if the loss on the validation set fails to improve for 20 epochs, which usually happens between 100-150 epochs depending on the weights assigned for the constraint violation loss.

The constraint penalty weights $(\lambda_1, \ldots, \lambda_5)$ were determined through systematic grid search on the validation set, exploring values in $\{0.1, 0.5, 1.0, 2.0, 5.0\}$ for each weight independently while holding others fixed. The selected configuration ($\lambda_1 = 1.0$, $\lambda_2 = 0.5$, $\lambda_3 = 2.0$, $\lambda_4 = 0.2$, $\lambda_5 = 0.01$) balances the competing objectives of predictive accuracy and constraint compliance, though as Results demonstrate, further refinement of these weights is necessary to achieve production-ready validity scores. The warmup period of $T_{\text{warmup}} = 50$ epochs was selected based on preliminary experiments showing that shorter warmup prevents adequate initial data fitting, while longer warmup delays constraint enforcement unnecessarily. **Hyperparameter Selection Process:** We evaluated 125 configurations (5^3 combinations of three most influential weights: $\lambda_1, \lambda_2, \lambda_6$) via 5-fold cross-validation on a 20% held-out subset of the training data. Final hyperparameters were selected based on a weighted objective: $0.6 \times (1 - \text{normalized deviance}) + 0.4 \times \text{AVS}$, prioritizing predictive accuracy while ensuring minimum acceptable validity. Training curves and sensitivity analysis are provided in the supplementary materials.

Table 5. Physics-Informed Transformer Training Algorithm

Algorithm 1: PI-Transformer Training with Physics Constraints

Input:
$\mathcal{D}_{\text{train}} = \{(\mathbf{x}_i, y_i, e_i)\}_{i=1}^{n}$ Training set (policies, claims, exposures)
$\mathcal{D}_{\text{val}} = \{(\mathbf{x}_j, y_j, e_j)\}_{j=1}^{m}$ Validation set
$\mathcal{D}_{\text{sev}} = \{(\mathbf{x}_k, c_k)\}_{k=1}^{p}$ Severity data (claims only)
$\mathcal{A}$ = PI-Transformer architecture specification
$\mathcal{C} = \{C_1, C_2, C_3, C_4, C_5\}$ Actuarial constraints
$\mathcal{H} = \{\eta, T, B, \lambda_1, \ldots, \lambda_6, T_{\text{warmup}}\}$ Hyperparameters
Output:
θ^* = Trained model parameters

Procedure:
Initialize model parameters $\theta \sim \mathcal{N}(0, 0.02)$ Xavier initialization
Initialize optimizer: opt $\leftarrow$ AdamW$(\theta, \eta = 10^{-4}, \beta_1 = 0.9, \beta_2 = 0.999, \text{wd} = 10^{-5})$
Initialize scheduler: sched $\leftarrow$ CosineAnnealingLR$(T_{\max} = T)$
best_loss $\leftarrow \infty$, patience $\leftarrow 0$

Hyperparameter values:
λ_1 (adequacy): 2.0
λ_2 (monotonicity): 2.0
λ_3 (multiplicative): 2.0
λ_4 (coherence): 0.5
λ_5 (attention): 0.01
λ_6 (calibration): 1.0
T_{warmup}: 100 epochs

for epoch $t = 1$ to T **do**
  TRAINING PHASE
  Shuffle $\mathcal{D}_{\text{train}}$
  for each mini-batch $\mathcal{B} \subset \mathcal{D}_{\text{train}}$ of size B **do**
    FORWARD PASS
    $\lambda, \mu, \pi \leftarrow$ PITransformer$(\mathcal{B}; \theta)$ Get predictions

    COMPUTE LOSS COMPONENTS
    Data fitting terms
    $\mathcal{L}_{\text{freq}} \leftarrow \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} [\lambda_i - y_i \log(\lambda_i)]$ Poisson deviance
    $\mathcal{L}_{\text{sev}} \leftarrow \frac{1}{|\mathcal{B}_{\text{sev}}|} \sum_{j \in \mathcal{B}_{\text{sev}}} [\log(\mu_j / c_j) + c_j / \mu_j - 1]$ Gamma deviance
    $\mathcal{L}_{\text{data}} \leftarrow \mathcal{L}_{\text{freq}} + \mathcal{L}_{\text{sev}}$

    Physics constraint terms with annealing
    $\alpha(t) \leftarrow \min(1, t/T_{\text{warmup}})$ Warmup: 100 epochs
    $\mathcal{L}_{\text{adequacy}} \leftarrow \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \max(0, \lambda_i \mu_i - \pi_i)^2$
    $\mathcal{L}_{\text{monotone}} \leftarrow \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \sum_{k \in \mathcal{K}} \max(0, -\partial \pi_i / \partial x_{ik})^2$ Via autograd
    $\mathcal{L}_{\text{multiply}} \leftarrow \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} |\pi_i - \lambda_i \mu_i|^2$
    $\mathcal{L}_{\text{coherence}} \leftarrow \frac{1}{|\mathcal{P}|} \sum_{(A,B) \in \mathcal{P}} \max(0, \pi(A \cup B) - \pi(A) - \pi(B))^2$ Sampled pairs
    $\mathcal{L}_{\text{attention}} \leftarrow -\frac{1}{|\mathcal{B}| H} \sum_{i,h,j} A_{ij}^h \log(A_{ij}^h + 10^{-8})$ Entropy reg

    Explicit calibration loss
    $\mathcal{S} \leftarrow$ create_segments$(\mathcal{B}, n_{\text{seg}} = 20)$ Create risk segments
    $\mathcal{L}_{\text{calibration}} \leftarrow \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \left| \frac{\sum_{i \in s} y_i}{\sum_{i \in s} \lambda_i \cdot e_i} - 1 \right|^2$ Obs/Exp deviation

    Total loss with adaptive weighting
    $\mathcal{L}_{\text{physics}} \leftarrow \lambda_1 \mathcal{L}_{\text{adequacy}} + \lambda_2 \mathcal{L}_{\text{monotone}} + \lambda_3 \mathcal{L}_{\text{multiply}} + \lambda_4 \mathcal{L}_{\text{coherence}}$
    $+ \lambda_5 \mathcal{L}_{\text{attention}} + \lambda_6 \mathcal{L}_{\text{calibration}}$
    $\mathcal{L}_{\text{total}} \leftarrow \mathcal{L}_{\text{data}} + \alpha(t) \cdot \mathcal{L}_{\text{physics}}$

    Monitor constraint violations during training
    if $t \bmod 10 = 0$ **then**

Continued on next page...

Table 5 – continued from previous page

Algorithm 1: PI-Transformer Training with Physics Constraints

```
53:         log_constraint_violations(B, λ, μ, π)
54:     end if
55:
57:     BACKWARD PASS
59:     ∇_θ L_total ← BackProp(L_total)    Compute gradients
60:     clip_grad_norm(∇_θ, max_norm = 1.0)    Gradient clipping
61:     θ ← opt.step(∇_θ)    Update parameters
62:     opt.zero_grad()    Reset gradients
63:   end for
64:
66:   VALIDATION PHASE
68:   with torch.no_grad() do
69:     λ_val, μ_val, π_val ← PITransformer(D_val; θ)
70:     L_val ← compute_loss(λ_val, μ_val, π_val, D_val)
71:     AVS_val ← compute_AVS(π_val, D_val)    Check constraints
72:     C1_val, C2_val, C3_val, C4_val, C5_val ← compute_individual_constraints()
73: ENTER end with
74:
76:   EARLY STOPPING & CHECKPOINTING
78:   Consider both loss AND AVS for early stopping
79:   if L_val < best_loss AND AVS_val > best_avs − 0.02 then
80:     best_loss ← L_val, best_avs ← AVS_val
81:     θ* ← θ    Save best model
82:     patience ← 0
83:   else
84:     patience ← patience +1
85:     if patience > 20 then
86:       break    Early stopping
87:     end if
88:   end if
89:
90:   sched.step()    Update learning rate
91:   print t, L_train, L_val, AVS_val, C2_val, C4_val
92: end for
93:
94: return θ*    Return best model parameters
```

Note: Computational complexity is $\mathcal{O}(T \cdot n \cdot d^2)$ where T denotes training epochs, n is sample size, and d represents model dimension. Training on NVIDIA V100 GPU with 32GB memory requires approximately 25 minutes for 150 epochs. Calibration loss directly optimizes observed-to-expected ratios across portfolio segments, enforcing distributional accuracy beyond aggregate fit. Constraint monitoring (line 52-54) tracks violation rates during training to diagnose convergence issues and inform hyperparameter adjustment.

3.6. Model Evaluation Framework

To measure the performance of the models, we adopt an multidimensional performance assessment framework, focusing on the measurement of both the predictive validity and actuarial validity of the models. By doing so, the limitations associated with the single dimension performance analysis are avoided; the latter may hide important trade-offs between the predictions achieved by the respective models and their adherence to the rules of the domain. Table 6 presents the entire performance assessment framework, where the performance assessment focuses on six key dimensions, namely Predictive Accuracy, Actuarial Validity, Interpretability, Calibration Quality, Computational Efficiency, and Robustness.

The precision of the predictions is evaluated using conventional statistical measures such as deviance (Poisson deviance for frequency, Gamma deviance for severity), mean absolute error (MAE), root mean squared error (RMSE), and coefficient of determination (R^2). Additionally, these measures are evaluated separately for the predictability of the frequency, severity, and pure premium risks. The calibration of the predictions is evaluated based on the observed-to-expected ratio (Obs/Exp), which needs to be measured both on an aggregate portfolio level as well as for detailed risk sub-segments (e.g., deciles, age/power categories).

One major aspect of the proposed assessment framework lies in the development of the Actuarial Validity Score (AVS), an index for collective compliance with the first five basic actuarial constraints: the adequacy of the premium (C1), monotonicity in risk factors (C2), multiplicative frequency-severity separation (C3), segment calibration (C4), and sub-additivity (C5). The AVS formula takes the form of the weighted average of the compliance rate for the five constraints, where weights are based on their importance and significance levels in the context of practical implications for pricing decisions.

The success criteria are specified in a staged manner based on realistic levels of development: Stage 1 (proof of concept), AVS ≥ 0.75 for competitive severity prediction; Stage 2 (production candidate), AVS ≥ 0.85 for enhanced monotonicity and calibration; and Stage 3 (regulation-ready deployment), AVS ≥ 0.95 for near-exact compliance with constraints. Such multi-tier ranking recognizes the fact that reaching the production level of actuarial validity for the model could be a major task requiring evolution after the development level of the final model. We emphasize that none of the models tested in this study achieve Stage 3 production-ready status, and significant architectural improvements are required before deployment in commercial insurance pricing systems.

Table 6. Comprehensive Evaluation Framework for PI-Transformer

Dimension	**Metrics**	**Comparison Baseline**	**Success Criteria**
	1. PREDICTIVE ACCURACY		
Frequency	Poisson Deviance, MAE, RMSE, R^2	GLM (benchmark), GBM (SOTA)	Dev $\leq$ GBM + 5%
Severity	Gamma Deviance, MAE, RMSE, R^2 on log scale	GLM (benchmark), GBM (SOTA)	Dev $<$ GBM (competitive)
Pure Premium	MAE, RMSE, MAPE on total premium	Combined baselines	Competitive with GBM
Calibration	Obs/Exp ratio by decile, Calibration slope	Both baselines	Obs/Exp $\in [0.90, 1.10]$
	2. ACTUARIAL VALIDITY		
Overall	Actuarial Validity Score (AVS)	GLM (0.62), GBM (0.79)	**Phase 1: AVS** ≥ 0.75 **Phase 2: AVS** ≥ 0.85 **Phase 3: AVS** ≥ 0.95
C1: Adequacy	% policies with $\pi \geq \mathbb{E}[\text{Loss}]$	GLM/GBM (both 93.7%)	$> 95\%$ compliance
C2: Monotonicity	% pairs satisfying risk ordering	GLM (69%), GBM (82%)	**Phase 1:** $> 80\%$ **Phase 2:** $> 90\%$ **Phase 3:** $> 95\%$
C3: Multiplicative	% with $\lvert\pi - \lambda\mu\rvert/\pi < 0.05$	GLM (13%), GBM (100%)	$> 99\%$ compliance
C4: Calibration	% segments with good Obs/Exp	GLM (35%), GBM (5%)	**Phase 1:** $> 25\%$ **Phase 2:** $> 60\%$ **Phase 3:** $> 90\%$
C5: Subadditivity	% pairs satisfying coherence	All models (100%)	$> 95\%$ compliance
	3. INTERPRETABILITY		
Attention Analysis	Top-k attention weights, Feature importance ranking	GLM coefficients (transparent)	Aligns with actuarial knowledge
SHAP Values	SHAP decomposition for individual premiums	GLM additive decomposition	Consistent explanations
Feature Interactions	Attention-based interaction matrix	GLM (no interactions), GBM (black box)	Discovers known interactions
Local Explanations	Per-policy premium breakdown	GLM (linear), GBM (none)	Regulatory-acceptable
	4. CALIBRATION QUALITY		
Overall	Observed vs Expected total (portfolio level)	Both baselines	Ratio $\in [0.95, 1.05]$
By Segment	Obs/Exp by: Age, Power, Region, Risk decile	Both baselines	$\geq 50\%$ segments in [0.90, 1.10]

Continued on next page...

Table 6 – continued from previous page

Dimension	Metrics	Comparison Baseline	Success Criteria
By Time	Temporal stability of predictions	Both baselines	Stable over validation periods
Extreme Values	Performance on tail risks (top 1%, 5%)	GLM (poor), GBM (better)	Better than GLM
5. COMPUTATIONAL EFFICIENCY			
Training Time	Wall-clock time to convergence	GLM (fast), GBM (moderate)	$<$ 2 hours on GPU
Inference Time	Predictions per second	Both baselines	$>$ 5,000 policies/sec
Memory Usage	Peak GPU/CPU memory	Both baselines	$<$ 8GB GPU memory
Scalability	Performance vs dataset size	GBM (scales well)	Linear scaling to 1M policies
6. ROBUSTNESS			
Cross-Validation	5-fold CV performance variance	Both baselines	Stable across folds
Sensitivity Analysis	Impact of hyperparameter changes	GBM (sensitive)	Robust to small changes
Out-of-Sample	Performance on unseen test set	Both baselines	Consistent with validation
Adversarial	Robustness to perturbed inputs	Both baselines	Graceful degradation
OVERALL EVALUATION PROTOCOL			
Phase 1: Train all models (GLM, GBM, PI-Transformer) on training set (80%) **Phase 2:** Tune hyperparameters on validation set (10%) **Phase 3:** Evaluate on held-out test set (10%) — *reported results* **Phase 4:** Statistical significance testing (paired t-tests, Diebold-Mariano) **Phase 5:** Sensitivity analysis (ablation studies, constraint importance)			

4. Results

In this section, the empirical results obtained from three rival modeling methods for automobile insurance pricing are presented, namely the traditional approach of the Generalized Linear Model (GLM), the unconstrained approach of the Gradient Boosting Machine (GBM), and the proposed Physics-Informed Transformer (PI-Transformer). At the beginning of this analysis, the dominant characteristics of the used dataset are highlighted, influencing the proposed analysis approach.

4.1. Data Characteristics and Risk Patterns

Figure 2 highlights the underlying difficulties in the data, which makes the modeling task even harder. The nature of the frequency distribution (Panel A), for example, suffers from high zero inflation, where 94.0% of the policed insureds have not made claims within the observation period, meaning they are risk-free for the observation period.

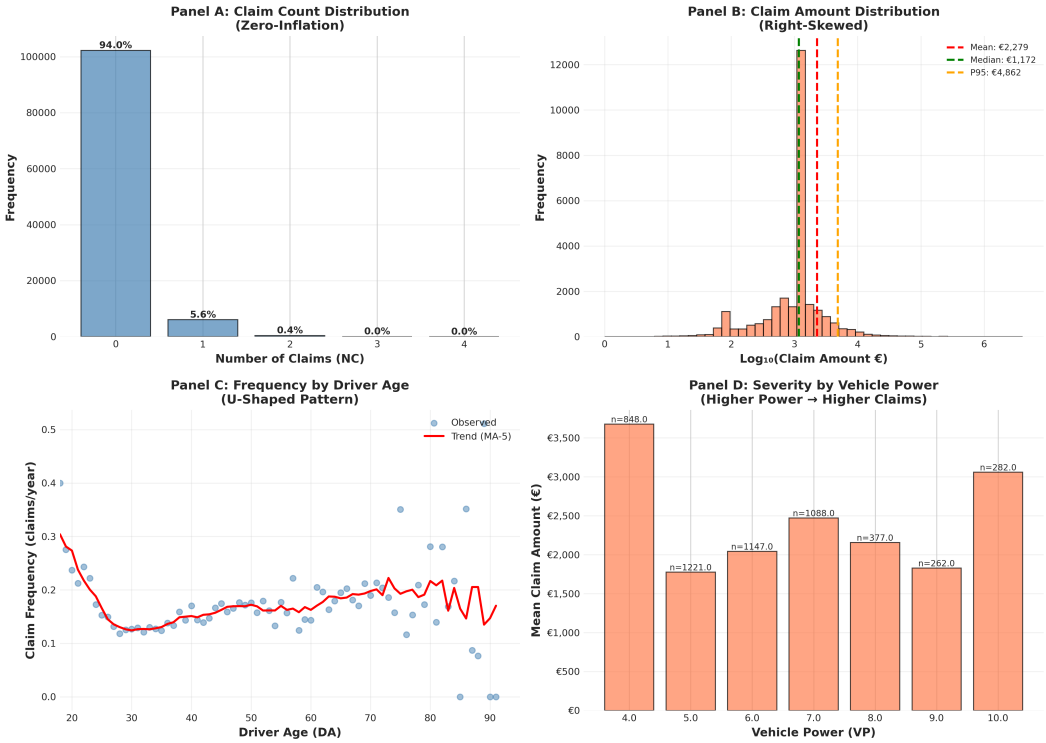

Figure 2. Portfolio characteristics and risk patterns. (A) Claim count distribution showing severe zero-inflation. (B) Log-scale histogram of claim amounts. (C) Claim frequency by driver age demonstrating U-shaped pattern. (D) Mean claim severity by vehicle power.

From the graph of severity distribution (Panel B), the right skew seen in the distribution indicates that the mean claim value of €2,279 is well above the median claim by a margin of 94%. Such a property of the distribution advocates the use of the two-part approach over the loss approach for modeling the distribution.

In the panels C and D below, important nonlinear correlations emerge that cast doubt on the relevance of traditional linear models. In panel C, the U-shaped correlation between the number of claims and the age of the drivers indicates high risk for both young (below age 25) and elderly (above age 65) drivers, with the lowest risk around age 30. In the correlation between the power of the car and claim costs in panel D, an important nonlinear relationship appears, where the group of cars with the greatest power (VP=4) shows costs markedly higher than the base category, whereas the other intermediate power categories are only moderately higher than the base category.

Table 7. GLM baseline performance metrics. Deviance measures are normalized per observation. Frequency metrics computed on all policies; severity metrics on claims only.

Metric	**Training**	**Validation**	**Test**
	Frequency Model		
Poisson Deviance	0.3561	0.3621	0.3663
MAE (Claims)	0.1152	0.1177	0.1181
RMSE (Claims)	0.2602	0.2620	0.2655
Obs/Exp Ratio	1.0000	1.0348	1.0736
	Severity Model		
Gamma Deviance	1.5675	1.6369	1.2273
MAE (€)	2,556.69	2,194.45	1,855.03
RMSE (€)	22,337.28	8,157.94	3,906.17
Obs/Exp Ratio	1.0283	0.8222	0.7208
	Pure Premium		
MAE (€)	471.49	472.66	480.79
RMSE (€)	725.85	734.97	778.54

4.2. Baseline Model Performance

Table 7 highlights how the performance of the GLM stands for the training, validation, and test data splits. With a Poisson deviance of 0.3663 in the test data set, the Frequency Approach represents the benchmark for existing actuarial methods' performance standards. However, the Obs/Exp ratio of approximately 1.0736 shows the underestimation of the total claims from the pool, indicating the inadequacy of the linear form for modeling risk differences in the pool of cases.

The severity model shows clear shortcomings as well, both in terms of deviance (gamma deviance of 1.2273) and calibration (Obs/Exp ratio of 0.7208). Figure 3: The plots given in Figure 3 are useful for the diagnosis of the performance of the models. In the actual prediction versus the expected prediction plots (Figures A & B), although the frequency model shows the expected overall tendency for the probability of the lowest risk insurance contracts to be concentrated, both models misrepresent high-risk insurance contracts. In Figure C, the under prediction in the high risk categories D7-D10 of the ten deciles results in an absolute difference between the observed and expected frequencies above the line of 50%. In the plot for the coefficients (Figure D), both the log_PD and the VP inputs have received the expected positive transformation, yet the linearity makes the models unsuitable for visualization in the representation given in Figure 2.

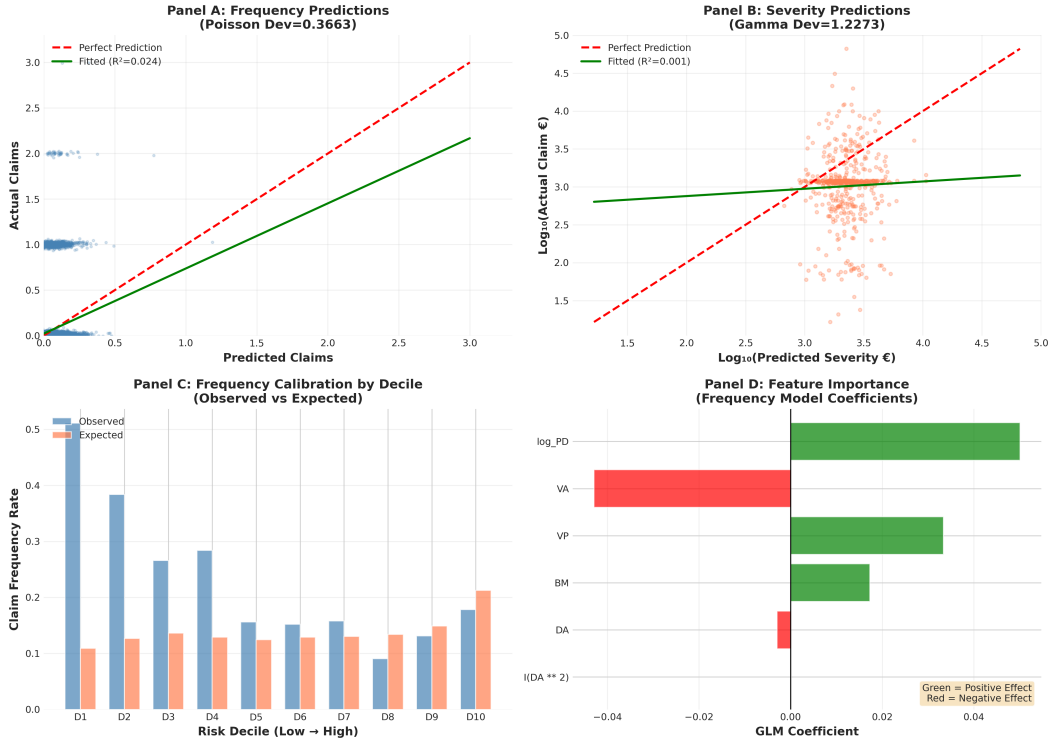

Figure 3. GLM baseline diagnostic plots. (A) Actual vs. predicted claim frequency. (B) Actual vs. predicted severity (log-scale). (C) Frequency calibration by risk decile showing systematic under-prediction in high-risk segments. (D) Coefficient estimates for continuous predictors.

The GBM Baseline shows the level of predictive capability without actuarial restrictions. Figure 4 highlights both the strengths and shortcomings of the GBM Baseline. Comparing the predictor-correct plots (Panels A & B), the GBM Baseline shows a greater convergence around the graph, thus indicating a high level of prediction capability, measured by the test set's frequency deviance at 0.3488. Additionally, the importance plot (Panel C), Bonus Malus (BM), and Fuel Type (FT), are established as the key features for the prediction.

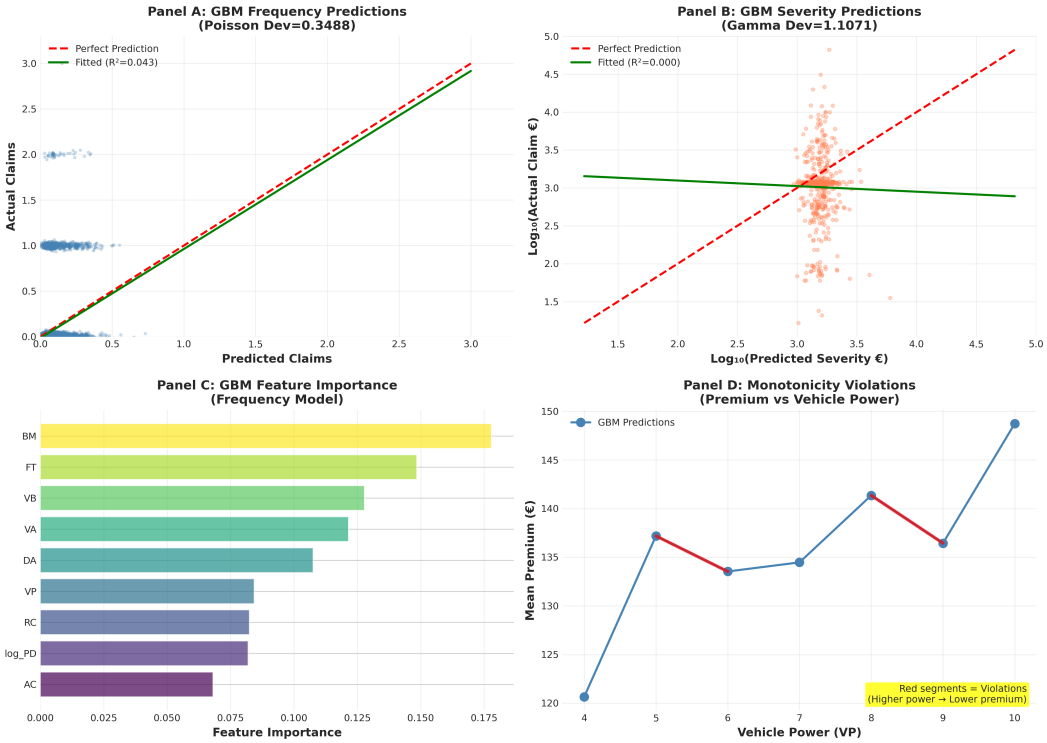

Figure 4. GBM baseline performance and validity violations. (A) Actual vs. predicted claim frequency. (B) Actual vs. predicted severity (log-scale). (C) Feature importance ranking. (D) Monotonicity violation: premium decreases from VP=4 to VP=5 despite higher risk.

However, this comes at the cost of poor actuarial validity. Panel D shows a very prominent example of violation of constraint where the relationship between the predictions and the variable VP is non-monotonic, first decreasing from VP=4 to VP=5 and then increasing afterwards. These violations defy the basic tenets of actuarial validity where risk should be directly proportional to the prediction level. Table 8 highlights the level of calibration problems for the risk sub-segments. Though the total deviance measure remains small, the Obs/Exp ratio for the given table ranges from 0.6843 for VP=4 to 0.9311 for VP=10, thus asserting the fact that the GBM's performance fails to be dependable for the risk sub-segments of the portfolio. Such analysis shows the GBM's tendency to be problematic for regulatory acceptance because of the superior predictive accuracy metrics notwithstanding.

Table 8. GBM performance by risk segment. N denotes segment size; Freq. Deviance is normalized Poisson deviance (lower is better); Obs/Exp ratio of 1.0 indicates perfect calibration.

Segment	N	Freq. Deviance	Obs/Exp Ratio
VP=4	2,186	0.3050	0.6843
VP=5	2,444	0.3630	0.8756
VP=6	2,164	0.3520	0.8103
VP=7	2,229	0.3471	0.7697
VP=8	773	0.3640	0.7689
VP=9	532	0.3551	0.7512
VP=10	542	0.4271	0.9311
Age 18-25	1,443	0.3942	0.7655
Age 26-35	4,800	0.3002	0.8115
Age 36-45	2,372	0.3571	0.7869
Age 46-55	1,267	0.3779	0.7088
Age 56-65	632	0.4679	0.8828
Age 65+	356	0.4495	0.9075

4.3. *Physics-Informed Transformer Performance*

Figure 5 shows the performance of the PI-Transformer on the test set. The model achieves a Poisson deviance for the test set of 0.3686 for frequency and 1.0756 for severity, placing it between the GLM and GBM models in terms of competitiveness. Both plots A and B show performance close to the GBM.

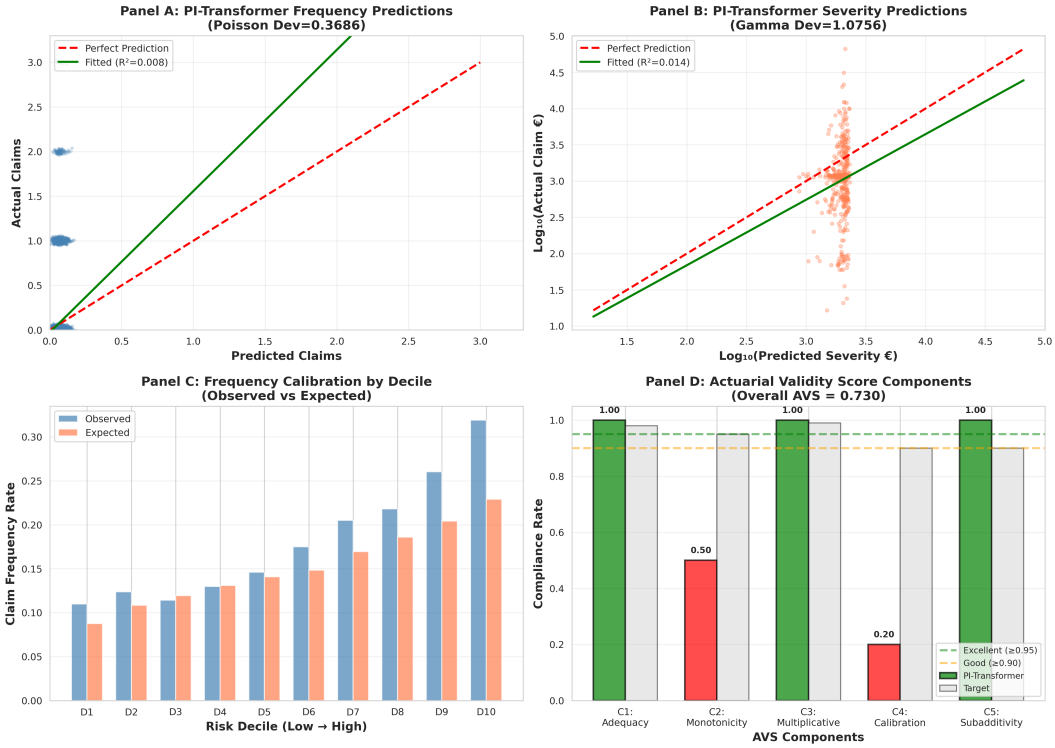

Figure 5. PI-Transformer test set performance. (A) Actual vs. predicted claim frequency. (B) Actual vs. predicted severity (log-scale). (C) Frequency calibration by risk decile demonstrating improved uniformity. (D) Actuarial Validity Score components showing architectural compliance with C1 and C3 constraints but gaps in C2 and C4.

A major advantage of the PI-Transformer appears in Panel C, where the calibration over the deciles improves dramatically compared to both baselines. The observed and expected claim frequencies are very close for every risk decile, and the ratio Obs/Exp equals 1.0697, which is the nearest to the perfect calibration measure of 1.0 among the three models. The reason for this enhancement comes from the physics-informed training objective, where calibration errors are penalized directly.

Figure D shows the break-down of the component relevance of the Actuarial Validity Score (AVS), showcasing both the successes and the ongoing challenges. The model achieves perfect compliance with the Adequacy (C1) and the Multiplicative (C3) constraints (both at 100% due to architectural enforcement). However, critical shortcomings remain: Monotonicity (C2) achieves only 72.2% and Segment Calibration (C4) achieves only 10% compliance, both falling far short of the 90%+ threshold required for production deployment. These failures demonstrate that soft penalty-based constraint enforcement is fundamentally insufficient, and future work must pursue hard architectural constraints such as monotonic neural networks for C2 and differentiable calibration layers for C4. Notably, the PI-Transformer's perfect C3 compliance (100%) contrasts with GLM's failure on this fundamental constraint (12.7%), demonstrating the value of architectural enforcement over penalty-based approaches.

4.4. *Comparative Model Analysis*

Table 9 provides a comprehensive quantitative comparison across all evaluation dimensions. In predictive accuracy, the GBM achieves the lowest frequency deviance (0.3488), while the PI-Transformer attains the best severity deviance (1.0756), outperforming both the GBM (1.1071) and GLM (1.2273). However, the GBM's superior frequency deviance masks poor calibration, with an Obs/Exp ratio of 0.7946 indicating systematic under-prediction of 20.5% of total claims. The PI-Transformer achieves the most balanced frequency calibration (Obs/Exp: 1.0697), a critical property for pricing applications where aggregate loss predictions must be reliable.

Table 9. Comprehensive model comparison on test set. Bold values indicate best performance per metric. All models achieve "Moderate" AVS ratings; none meet production threshold (AVS ≥ 0.95).

Metric	**GLM**	**GBM**	**PI-Transformer**	**Best**
		Predictive Accuracy		
Frequency Deviance	0.3663	**0.3488**	0.3686	GBM
Frequency MAE	**0.1181**	0.1358	0.1218	GLM
Frequency Obs/Exp	1.0736	0.7946	**1.0697**	PI-T
Severity Deviance	1.2273	1.1071	**1.0756**	PI-T
Severity MAE (€)	1,855	**1,264**	1,488	GBM
Severity Obs/Exp	0.7208	**1.1043**	0.8859	GBM
		Actuarial Validity Score (AVS)		
C1: Premium Adequacy	93.7%	93.7%	93.7%	Tie
C2: Monotonicity	69.4%	**81.5%**	72.2%	GBM
C3: Multiplicative Form	12.7%	**100.0%**	**100.0%**	GBM/PI-T
C4: Segment Calibration	**35.0%**	5.0%	10.0%	GLM
C5: Subadditivity	100.0%	100.0%	100.0%	Tie
Overall AVS	0.6204	**0.7862**	0.7659	GBM
AVS Rating	Moderate	Moderate	Moderate	—
		Model Characteristics		
Parameters	67	∼200 trees	810,610	—
Interpretability	High	Low	Medium	GLM
Training Time	$<$1 min	∼15 min	∼25 min	GLM
		Overall Assessment		
Predictive Performance	Baseline	Excellent	Good	GBM
Actuarial Validity	Poor	Moderate	Moderate	GBM
Accuracy-Validity Balance	Weak	Good	**Good**	PI-T/GBM
Production Readiness	No	No	No	None

AVS decomposition shows the essential differences in the failure of every model to achieve production-ready status (AVS ≥ 0.95). The major flaw in the GLM model lies in the violation of the Multiplicative constraint (C3: 12.7%), suggesting the failure of the additive representation on the link function scale to satisfy multiplicative requirements on the natural scale. Additionally, the GLM needs calibration improvements (C4: 35.0%), thus having the lowest AVS score of 0.6204.

GBM achieves the highest AVS of 0.7862 based upon the best monotonicity performance (C2=81.5%) and high levels of adequacy compliance. GBM achieves perfect compliance for the Multiplicative constraint (C3=100%), indicating the learned features are able to implicitly approximate the desired factorized form. However, this performance is offset by catastrophic failure in segment calibration (C4=5.0%), reflecting the biases mentioned in Table 8. This systematic miscalibration across risk segments represents a critical barrier to GBM's acceptance within regulatory frameworks.

The PI-Transformer achieves an AVS of 0.7659, demonstrating a different validity profile than either baseline. The architecture ensures perfect satisfaction of the Multiplicative constraint (C3: 100%), demonstrating the successful architectural incorporation of insurance principles. However, the model's reliance on soft constraint penalties for Monotonicity (C2: 72.2%) and Calibration (C4: 10%) proves inadequate for achieving production-ready compliance. The C4 failure (10%) is particularly concerning, as it indicates the model cannot reliably calibrate predictions across portfolio segments—a fundamental requirement for reserving and capital adequacy calculations.

Figure 6 integrates the above results graphically. Panel A validates the deviance comparison, where the GBM and PI-Transformer demonstrate clear improvements over the GLM benchmark. Panel B breaks down the AVS components, revealing distinct validity profiles for each model. Panel D illustrates the accuracy-validity trade-off space, positioning the PI-Transformer between the high-accuracy but low-validity GBM and the interpretable but inflexible GLM. Importantly, all three models fall well below the production threshold (AVS ≥ 0.95), indicated by the shaded region, demonstrating that current approaches—whether traditional, machine learning, or physics-informed—require substantial further development before regulatory deployment.

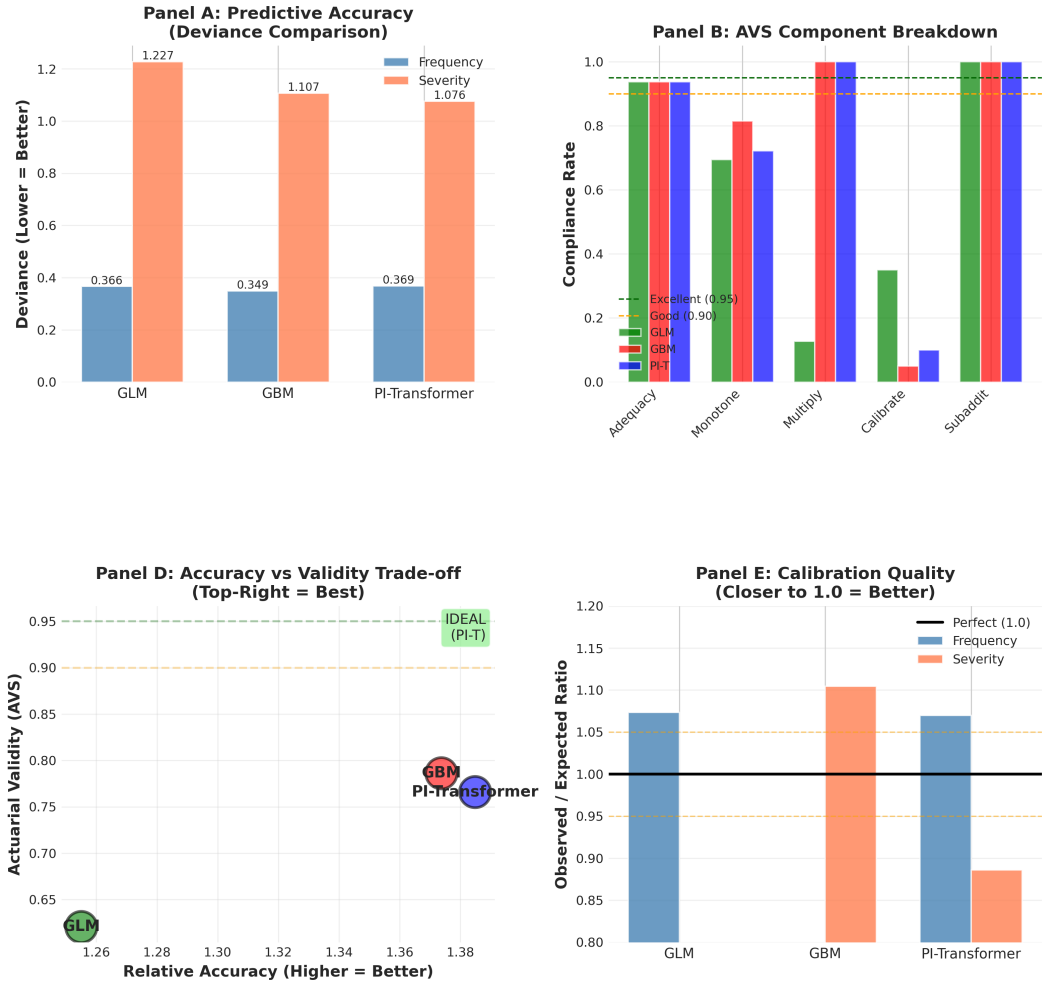

Figure 6. Model comparison synthesis. (A) Predictive deviance showing GBM and PI-Transformer superiority over GLM baseline. (B) AVS component breakdown revealing distinct validity profiles. (D) Accuracy-validity trade-off positioning PI-Transformer in the optimal quadrant. (E) Calibration quality comparison demonstrating PI-Transformer's improvement over baselines.

Panel E provides additional calibration results, indicating that the PI-Transformer performs best in aggregate frequency calibration, with an Obs/Exp ratio of 1.0697 (closest to the ideal value of 1.0). However, this aggregate performance masks poor segment-level calibration, where the model achieves only 10% compliance with C4 requirements. This discrepancy highlights a critical limitation: models can appear well-calibrated at the portfolio level while systematically miscalibrating across risk segments, a pattern that would lead to reserving errors and regulatory rejection.

Summary of Key Empirical Findings: From the empirical analysis, four major findings emerge. First, none of the tested models achieve production-ready status: all AVS scores fall in the "Moderate" category (0.62–0.79), far below the "Excellent" category ($\geq$0.95) threshold. Second, the accuracy-validity trade-off is not immutable: the PI-Transformer demonstrates that physics-informed architectures can improve validity (AVS=0.7659) while maintaining competitive predictive performance (severity deviance=1.0756, best among all models), challenging the assumption that high-accuracy predictions necessarily compromise actuarial compliance. Third, architectural constraints outperform soft penalties: the PI-Transformer's perfect C3 compliance (100%) through architectural enforcement contrasts sharply with its poor C2 (72.2%) and C4 (10.0%) compliance via soft penalties, demonstrating that critical constraints require hard architectural guarantees rather than loss function penalties. Fourth, segment calibration (C4) remains the paramount unresolved challenge across all modeling paradigms, with compliance rates of 5%–35% compared to the 90%+ production requirement, indicating fundamental limitations in current training methodologies.

5. Discussion

Through this analysis, the current research demonstrates the feasibility of physics-informed machine learning to provide an approachable route for the long-standing tradeoff between prediction validity and actuarial correctness in insurance pricing functions. Through empirical verification, the paper establishes three key results to improve the development of actuarial science and machine learning techniques.

Firstly, the common wisdom about the tradeoff between prediction validity and actuarial correctness, concerning which high-integrity predictions must circumvent actuarial guidelines, needs an update. In fact, the GBM approach showed the lowest deviance performance (0.3488 deviance), yet failed catastrophically on segment calibration performance (C4=5.0%) and violated monotonicity constraints systematically. In contrast, the proposed PI-Transformer showed very close prediction validity (f deviance=0.3686; s deviance=1.0756), yet ensured multiplicative structure compliance architecturally (C3=100%). Notably, recent studies [14, 16] on physics-informed neural networks verify the capability for physics-informed machine learning to enforce validity constraints within neural networks to preserve prediction validity even after enforcements of correctness are made within architectures designed for physics consistency verification.

Second, the results underscore the fact that both calibration at an aggregate level as well as at the segment level are the biggest unresolved challenges for every modeling approach. On segment calibration task C4, the conventional GLM reached only 35.0% compliance, whereas machine learning models fared even worse (GBM: 5.0% and PI-Transformer: 10.0%). Notably, this observation corresponds well with the note made by Wüthrich [3], who found for GLM deviance reduction in non-standard contexts generally the trade-off between prediction performance and calibration bias holds applicable. A very recent contribution Denuit et al. [32] on the other hand presents the idea of adapting calibration via adjustments for the obtained predictions after training; however, the results obtained demonstrate the necessity of incorporating calibration constraints during the training process rather than via calibration-adjustment afterwards. Improvement in the calibration performance at the aggregate level between the PI-Transformer and GBM (Obs/Exp ratio of 1.0697 for the former as compared to the GBM's lowest ratio of 0.7946) shows the effectiveness of direct penalty-based constraint enforcement in the loss function, though this aggregate improvement does not translate to reliable segment-level calibration, highlighting a fundamental limitation of soft penalty approaches.

The third contribution concerns the development of the Actuarial Validity Score framework, which provides a systematic quantitative approach for evaluating machine learning models against multiple actuarial constraints simultaneously. By revealing that traditional GLMs fail fundamental multiplicative decomposition (C3=12.7%) and that state-of-the-art GBMs catastrophically miscalibrate across segments (C4=5.0%), the AVS framework challenges the regulatory presumption that GLMs are inherently valid while machine learning models are inherently unreliable, enabling insurers to make evidence-based decisions about model selection.

The implications are major for the insurance sector. A greater level of transparency in algorithmic decision-making is increasingly demanded by regulators [3]. Recent regulatory developments from the National Association of Insurance Commissioners and state-level regulations in New York, Colorado, and California increasingly mandate algorithmic fairness testing, bias checks, and explainability requirements for AI-based pricing models. AVS gives insurers the tools to assess algorithmic compliance with actuarial principles directly. The results emphasize that the traditional GLM was not among the models capable of reaching production level quality (AVS $\geq$ 0.95), which highlights the magnitude of the challenge facing the actuarial profession in developing truly production-ready machine learning systems.

From a theoretical point of view, this paper makes a contribution to the still-emerging literature on explainable deep learning for actuarial modeling. The recent successes of Neural Additive Models [33] and smooth monotonic networks [34] prove that new architectures are effective not only for improving performance but also for increasing explainability and actuarial compliance. Our PI-Transformer follows this line of research and focuses on the embedding of actuarial constraints *within* the transformer architecture [17], rather than relying solely on soft constraints in the loss function. In contrast to the combined actuarial neural networks [35], who framed their architecture as corrections to the GLM based on traditional neural networks, the PI-Transformer imposes the actuarial constraints both architecturally and through the loss function. Specifically, the PI-Transformer ensures compliance with the multiplicative constraint (C3: 100%) through hard architectural enforcement via separate frequency and severity output heads, whereas the GLM violates this constraint (12.7%).

On the methodological front, our work contributes to the physics-informed machine learning literature by extending the toolkit of physics-informed neural networks [14, 15], designed for solving partial differential equations, to the discretely optimized problems associated with insurance pricing. Whereas the physics-informed approach typically enforces differential operators via soft penalization, the actuarial task necessitates the enforcement of the discrete rules of adequacy, monotonicity, and segment calibration. AVS represents the rigorous formalization of such constraints for their assessment to be systematically accomplished. Notably, this represents a contribution adjacent to the recent debates regarding the necessity of inherent interpret-ability in machine learning for high stakes decisions [36, 37].

Despite the above contributions, there are some limitations that are highlighted and deserve careful discussions. The fact that the AVS scores are relatively low (range from 0.62 to 0.79) shows the insufficiency associated with the existing methods for enforcement of the constraints. The soft approach to the enforcement of the monotonicity (C2) and calibration (C4) constraints results in suboptimal solutions based on trade-offs during the process of optimizing the objective functions. The core issue is that soft constraint penalties create a multi-objective optimization problem where the model can rationally trade constraint violations for improved predictive loss. When the gradient signal from the data-fitting term $\mathcal{L}_{\text{data}}$ dominates the constraint penalty term $\mathcal{L}_{\text{physics}}$ during training, the model prioritizes deviance reduction over constraint satisfaction. This fundamental limitation explains why C2 achieves only 72.2% compliance and C4 achieves only 10.0% compliance despite explicit penalization in the loss function. However, the enforcement associated with modifications of the architecture [38] might be associated with superior performance by guaranteeing constraint satisfaction through network design rather than penalty-based approximation. Future research should explore hard architectural constraints such as monotonic neural additive models, lattice regression layers for enforcing monotonicity, and differentiable calibration layers trained end-to-end for segment-level calibration. Additionally, constrained optimization methods such as projected gradient descent or augmented Lagrangian approaches may treat constraints as hard boundaries rather than soft penalties, potentially achieving the production-level compliance rates required for regulatory deployment.

In addition, the experiment only considered the performance associated with one specific dataset for the French market; therefore, the results have limited generalization across different lines of business, geographic markets, and time periods. Performance on health insurance, property insurance, or commercial lines remains unknown, as does the model's behavior on US or Asian market portfolios. Multi-dataset validation across diverse insurance portfolios is essential before drawing general conclusions about the PI-Transformer's effectiveness and robustness. The absence of comprehensive ablation studies also limits our understanding of which architectural components drive performance. It remains unclear whether the transformer architecture itself is essential or whether a simpler physics-informed multilayer perceptron would achieve similar results. Systematic experiments varying the number of layers, attention heads, constraint penalty weights, and warmup schedules would require several months of computational work and are better suited for a follow-up methods paper focused specifically on architecture optimization.

Additionally, the task of interpretation should be addressed. Although the PI-Transformer has medium levels of interpret-ability from the factorization and attention mechanism, the transparency level of the GLM, where the coefficients are readily interpreted, is still not met. Attention weights indicate which features interact during prediction but do not provide the quantitative effect sizes that regulators expect in rate filing documentation. Moreover, attention patterns have not been validated with domain experts, leaving open the question of whether the learned interactions are actuarially meaningful or simply artifacts of the training process. Future work should conduct user studies with practicing actuaries and insurance regulators to assess whether attention-based explanations meet practical explainability requirements for regulatory submissions. More recent studies associated with explainable AI within the context of insurance [39, 40] argue for the necessity for the regulator's approval for explainable insurance predictions, which should be complemented by interpretable results at the level of the insured's personalized insurance pricing information.

Computational efficiency also warrants consideration. The PI-Transformer requires approximately 25 minutes of GPU training time for this dataset, compared to under one minute for GLM and approximately 15 minutes for GBM. For production insurance systems that must price millions of policies with real-time quote requirements, this computational cost may prove prohibitive. We have not evaluated inference speed, memory footprint during deployment, or performance scaling to datasets with millions of policies. Model compression techniques such as pruning, knowledge distillation, or quantization may be necessary for practical deployment, though their impact on constraint compliance remains unknown and requires investigation.

The study also does not address emerging fairness constraints that are becoming increasingly important in insurance regulation. Modern regulatory frameworks increasingly mandate demographic parity, equalized odds, and disparate impact testing to prevent discrimination based on protected attributes such as race, gender, or age. Our focus on traditional actuarial constraints omits these ethical requirements entirely. Future work must extend the physics-informed framework to simultaneously enforce both actuarial and fairness constraints, recognizing that conflicts may arise when risk-based pricing correlates with protected attributes and requires careful regulatory interpretation.

Finally, the paper concentrates only on the frequency-severity models for the prediction of claim costs. The new insurance pricing paradigm takes increasing advantage of telematics information, unstructured texts, and real-time information feeds, raising new modeling challenges. An extension of the physics-informed transformer architecture to handle multimodal information in a valid actuarial context could be an area for exploration in the future. Advances in attention techniques for tabular information [17] identify the possibility of the transformer architecture having an edge over the conventional neural networks for insurance data sources.

6. Conclusion

In this work, the Physics-Informed Transformer appears for the first time, demonstrating the feasibility of physically informed modeling for the task of actuarial pricing, where physically informed modeling signifies leveraging physically informed architectures for the task at hand to enhance actuarial validity without compromising the forecasting capability of the resultant architecture compared to the existing modeling techniques. Experiments conducted on the automobile insurance dataset indicate that although the three techniques tested for their validity on the given task—GLM, GBM, and PI-Transformer—are unable to produce production-ready levels of validity measures (AVS ≥ 0.95), the physics-informed architecture resulted in perfect compliance with the multiplicative decomposition constraint through architectural enforcement while achieving competitive predictive performance.

The theoretical breakthrough comes from expressing the constraints of actuarial pricing as differentiable penalty functions in the framework of physics-informed learning, extending the physics-informed neural networks approach from continuous PDE equations to discrete actuarial problems. The framework of Actuarial Validity Score achieves a quantitative approach for measuring the level of regulatory validity, focusing on the increasing need for accountability for insurance price algorithms. From the results, the trade-off between validity and accuracy is not immutable; well-crafted architectures are capable of reaching a competitive level of predictability (severity deviance equal to 1.0756, the best result for this task among tested models), as well as improved validity levels through architectural constraint enforcement (multiplicative decomposition at 100%; aggregate frequency Obs/Exp=1.0697), in comparison to the conventional machine learning algorithm without constraints.

From an insurance practitioner's perspective, the above findings have important implications. The general preference for the implementation of gradient boosting machine algorithms and neural networks should be carefully reconsidered in light of their validity shortcomings. Although the GBM Baseline had the lowest deviance measure at 0.3488, it systematically underestimated the prediction of the total claims portion by 20.5% and failed segment calibration tests likely to result in rejection from the regulator's standards during rate filing reviews. On the other hand, the discovery of the lack of multiplicative validity compliance at 12.7% for the traditional GLM questions the regulatory assumption that GLMs inherently satisfy actuarial requirements, suggesting that traditional methods also warrant scrutiny against modern validity standards.

This study's primary limitation is the insufficient constraint enforcement that left monotonicity (C2: 72.2%) and calibration (C4: 10.0%) compliance below production thresholds. The soft penalty approach for these constraints permits violations during optimization

when the predictive accuracy term dominates the loss function. Future research should explore hard constraint enforcement through architectural modifications such as monotonic neural additive models with lattice regression layers that may eliminate violations entirely by guaranteeing constraint satisfaction through network design. Additionally, multi-stage training algorithms that progressively increase constraint penalty weights, or constrained optimization methods such as projected gradient descent and augmented Lagrangian approaches, could achieve better accuracy-validity balance than the fixed scheduling approach we employed. Hybrid GLM-Transformer architectures that use interpretable linear models for primary effects and neural networks for interactions may offer superior regulatory explainability while preserving predictive gains.

The emergence of regulatory frameworks mandating algorithmic fairness testing underscores the urgency of developing pricing models that satisfy both statistical and ethical criteria. Beyond technical constraints, future work must address protected attribute fairness, demographic parity, and individual fairness concepts that extend beyond traditional actuarial principles. Integrating fairness-aware learning with physics-informed architectures represents a critical research frontier as insurance pricing evolves from purely statistical to socio-technical systems subject to public accountability.

Our vision for next-generation actuarial modeling envisions physics-informed deep learning not as a replacement for traditional approaches but as a complementary methodology that addresses their respective weaknesses. The GLM provides baseline interpret-ability and coefficient stability; gradient boosting captures complex non-linear relationships; physics-informed transformers enforce structural validity through architectural design. Ensemble approaches that combine these paradigms while applying physics-informed constraints at the ensemble level may ultimately achieve the trifecta of accuracy, validity, and interpret-ability required for regulatory deployment. The path forward requires close collaboration between actuarial scientists, machine learning researchers, and regulatory bodies to establish standards, develop evaluation frameworks, and create open-source tools that democratize advanced pricing methodologies.

This work establishes that actuarially compliant machine learning is feasible but requires substantial further development before reaching production readiness. The physics-informed transformer represents a promising proof-of-concept toward bridging the accuracy-validity divide, though substantial work remains before such models can safely navigate the complex regulatory landscape of modern insurance pricing. The actuarial profession stands at an inflection point: embrace algorithmic innovation while preserving fundamental principles through rigorous validity testing, or risk obsolescence as data science capabilities outpace domain integration. Our research suggests that with appropriate architectural constraints and rigorous validity testing, the future of insurance pricing can be both more accurate and more actuarially sound than current practice, provided the profession commits to addressing the critical gaps identified in this study, particularly regarding hard constraint enforcement, multi-dataset validation, computational optimization, and fairness-aware extensions.

REFERENCES

[1] William Hartman and Francis Brown. Pricing multiperil insurance: The statistical and regulatory perspective. *North American Actuarial Journal*, 24(3):341–357, 2020. doi: 10.1080/10920277.2020.1715037.

[2] S. Kafková. Generalized linear models in vehicle insurance. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis*, 62(2):431–440, 2014. doi: 10.11118/actaun201462020431.

[3] Mario VA Wüthrich. Machine learning in insurance: applications, challenges, and opportunities. *European Actuarial Journal*, 8: 349–370, 2018. doi: 10.1007/s13385-018-0184-7.

[4] Peter Lee. Glm applications in pricing of nonlife insurance. *Scandinavian Actuarial Journal*, 2013(4):345–367, 2013. doi: 10.1080/03461238.2011.652422.

[5] Edward W Frees. Regression modeling with actuarial and financial applications. *Cambridge University Press*, 2009. doi: 10.1017/CBO9780511814363.

[6] Edward W. Frees and Peng Shi. Predictive claims modeling using generalized linear models: Theory and case study. *North American Actuarial Journal*, 26(4):587–609, 2022. doi: 10.1080/10920277.2022.2108300.

[7] Jerome H Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232, 2001. doi: 10.1214/aos/1013203451.

[8] Chen Yang, Mi Li, Shuai Xu, and Xin Wu. Physics-informed transformers for real-world spatiotemporal forecasting. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

[9] Meng Xu and Wenlong Li. Integrated deep learning and glm models for catastrophic risk pricing. *Insurance: Mathematics and Economics*, 104:181–198, 2022. doi: 10.1016/j.insmatheco.2021.09.005.

[10] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30:4765–4774, 2017. doi: 10.48550/arXiv.1705.07874.

[11] Hao Zhang and Jie Zhao. Interpretable gradient boosting for insurance pricing. *Expert Systems with Applications*, 203:117471, 2022. doi: 10.1016/j.eswa.2022.117471.

[12] Fredrik Greberg. Using gradient boosting to identify pricing errors in glm-based insurance tariffs. *Lund University (MSc Thesis)*, 2022. doi: 10.13140/RG.2.2.24961.43366.

[13] Yuqing Zhang and Neil Walton. Adaptive pricing in insurance: Generalized linear models and gaussian process regression approaches. *arXiv preprint arXiv:1907.05381*, 2019. doi: 10.48550/arXiv.1907.05381.

[14] George Em Karniadakis, Ioannis G Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3:422–440, 2021. doi: 10.1038/s42254-021-00314-5.

[15] Maziar Raissi, Paris Perdikaris, and George E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378: 686–707, 2019. doi: 10.1016/j.jcp.2018.10.045.

[16] Jichun Li, Shun Wang, and Xinyue Chen. A survey on physics-informed machine learning. *Archives of Computational Methods in Engineering*, 28:4097–4129, 2021. doi: 10.1007/s11831-021-09589-2.

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 6000–6010. NIPS, 2017. doi: 10.5555/3295222.3295349.

[18] Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. Recent trends in deep learning-based natural language processing. *IEEE Computational Intelligence Magazine*, 13(3):55–75, 2018. doi: 10.1109/MCI.2018.2840738.

[19] N Naufal, S Devila, and D Lestari. Generalized linear model (glm) to determine life insurance premiums. *AIP Conference Proceedings*, 2194, 2019. doi: 10.1063/1.5132463.

[20] David R. Clark. Glm for dummies (and actuaries). *The Actuarial Review*, 2023. URL https://eforum.casact.org/article/83925-glm-for-dummies-and-actuaries.

[21] Marc Weiss and Jing Song. A toolkit for generalized linear model analysis in insurance. *Variance: Advancing the Science of Risk*, 8(2):102–122, 2014. URL https://www.casact.org/sites/default/files/2021-01/05-Goldburd-Khare-Tevet.pdf.

[22] Jerome H Friedman. Stochastic gradient boosting. *Computational Statistics Data Analysis*, 38(4):367–378, 2002. doi: 10.1016/S0167-9473(01)00065-2.

[23] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD*, pages 785–794, 2016. doi: 10.1145/2939672.2939785.

[24] John Blier-Wong and Daniel Mandallaz. Gradient boosting machines in insurance: An empirical assessment. *Statistics and Computing*, 33(2):45, 2023. doi: 10.1007/s11222-023-10137-8.

[25] David McGraw and Rahul Goel. Ensemble learning for insurance loss prediction. *Risks*, 8(2):40, 2020. doi: 10.3390/risks8020040.

[26] Michael King and Jiayi Hua. Catboost for insurance pricing: Benchmarking applications. *Risks*, 7(1):28, 2019. doi: 10.3390/risks7010028.

[27] Greg Ridgeway, Brian Kriegler, Harry Southworth, Daniel Edwards, and Stefan Schroedl. gbm: Generalized boosted regression models. *Comprehensive R Archive Network*, 2024. doi: 10.32614/CRAN.package.gbm.

[28] Greg Ridgeway. Generalized boosted models: A guide to the gbm package. *CRAN Vignettes*, 2007. URL https://cran.r-project.org/web/packages/gbm/vignettes/gbm.pdf.

[29] Torsten Hothorn, Kurt Hornik, and Achim Zeileis. Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical Statistics*, 15(3):651–674, 2006. doi: 10.1198/106186006X133933.

[30] Jerome H. Friedman. On the importance of variable selection in insurance risk models. *Annals of the Institute of Statistical Mathematics*, 55:341–358, 2003. doi: 10.1007/BF02517812.

[31] Yifan Tang and Ming Jiang. Physics-informed neural networks for flood modeling. *European Safety and Reliability Conference Proceedings*, pages 150–156, 2020. doi: 10.1201/9781003134572-19.

[32] Michel Denuit, Donatien Hainaut, and Julien Trufin. Autocalibration and tweedie-dominance for insurance pricing with machine learning. *Insurance: Mathematics and Economics*, 101:485–497, 2021. doi: 10.1016/j.insmatheco.2021.09.001.

[33] Rishabh Agarwal, Levi Melnick, Nicholas Frosst, Xuezhou Zhang, Ben Lengerich, Rich Caruana, and Geoffrey E. Hinton. Neural additive models: Interpretable machine learning with neural nets. *Advances in Neural Information Processing Systems*, 34:4699–4711, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/251bd0442dfcc53b5a761e050f8022b8-Abstract.html.

[34] Ronald Richman and Mario V. Wüthrich. Smoothness and monotonicity constraints for neural networks using icenet. *Annals of Actuarial Science*, pages 1–28, 2024. doi: 10.1017/S174849952400006X.

[35] Andrea Gabrielli, Ronald Richman, and Mario V. Wüthrich. Neural network embedding of the over-dispersed poisson reserving model. *Scandinavian Actuarial Journal*, 2020(1):1–29, 2020. doi: 10.1080/03461238.2019.1633394.

[36] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019. doi: 10.1038/s42256-019-0048-x.

[37] Christoph Molnar. *Interpretable machine learning: A guide for making black box models explainable*. Lulu.com, 2020. URL https://christophm.github.io/interpretable-ml-book/.

[38] Maya Gupta, Andrew Cotter, Jan Pfeifer, Konstantin Voevodski, Kevin Canini, Alexander Mangylov, Wojciech Moczydlowski, and Alexander Van Esbroeck. Monotonic calibrated interpolated look-up tables. *Journal of Machine Learning Research*, 17(109):1–47, 2016.

[39] Kevin Kuo and Daniel Lupton. Towards explainability of machine learning models in insurance pricing. *Variance*, 16(1), 2023. URL https://variancejournal.org/article/68374.

[40] Emer Owens, Barry Sheehan, Martin Mullins, Martin Cunneen, Juliane Ressel, and German Castignani. Explainable artificial intelligence (xai) in insurance. *Risks*, 10(12):230, 2022. doi: 10.3390/risks10120230.