KPMG

Navigating AI model review

Insurance case studies

KPMG. Make the Difference.

January 2026

Context

As Artificial Intelligence (AI) adoption expands across financial services, insurers must evolve their model review approach to harness AI's power and benefits while managing risks such as bias, explainability, and drift.

A robust framework for AI model review is critical for ensuring regulatory trust, audit readiness, risk mitigation, and customer fairness.

Existing Model Risk Management (MRM) frameworks should be adapted to reflect AI-specific complexities such as biases in data-driven decision-making, model adaptability, and potential black-box behaviours.

AI model review: No one size fits all

The deployment of Artificial Intelligence (AI) spans a broad spectrum, with each AI model possessing unique characteristics that dictate the nature of its associated risks. Consequently, effective model review techniques must be tailored to the specific AI model type and how it is used. Below, we explore several prominent AI model types commonly used in the insurance sector, and the critical risks that demand careful consideration.

Types of AI models:

Anomaly detection models

Anomaly detection models leverage historical data alongside machine learning techniques to uncover patterns that facilitate predicting future events or identifying anomalies such as fraud or system malfunctions. One prominent use case for insurers is fraud detection, where these models highlight suspicious transactions to specialists for investigation.

Implementing these models involves considerable challenges, including risks of high instances of false positives or negatives, dependence on predefined thresholds and often limited transparency in explaining why specific events are flagged. Additionally, if the historical data used is biased, the model may reinforce or exacerbate these biases, raising concerns about fairness and potentially resulting in discriminatory consequences.

Generative AI (GenAI) models

GenAI models are trained on vast datasets, such as text, images, or data to produce new content in response to specific prompts. GenAI models are highly adaptable for a range of creative and analytical applications.
In the insurance sector, for example, GenAI can effectively summarize complex claim documents. However, utilising GenAI carries specific risks such as "hallucination," where the models generate factually inaccurate but plausible information and considerable variability in output which can affect reliability. Additional concerns include prompt injection attacks, where malicious inputs can alter the output and privacy issues, as the extensive training data may unintentionally include sensitive information.

Computer vision models

Computer vision models are designed to interpret and analyse visual data such as images and videos. These models perform functions including object detection, recognition, and segmentation. They typically require fine-tuning for specific tasks and demand significant computational resources.

In the insurance sector, a common application is assessing damage for claims, where the model evaluates images of damaged properties or vehicles to estimate repair costs.

However, limited generalisation remains as a major challenge. Models may underperform on visual data outside their training scope. The lack of explainability, which complicates understanding the reasoning behind their decisions is also a particular challenge. These issues underscore the necessity for a tailored approach to AI governance and model evaluation, as a universal method is inadequate given the complexity and rapid advancement of AI technologies.

The following section of this article will explore the key model review techniques, providing practical examples through use cases for the three AI model types mentioned above.

Use case 1 - Fraud detection model

AI model type: Anomaly detection model

An anomaly detection model is employed to identify potentially fraudulent claims by flagging them for further investigation according to a specified fraud probability threshold. This model analyses individual claims, utilizing historical claims data to recognize patterns associated with fraudulent activity. Any new claims that has exceeded the fraud probability threshold will be flagged to specialists for investigation.

A primary challenge involves fine-tuning the fraud probability threshold to achieve a balance between accurately identifying the frauds and adhering to operational limitations, such as the capacity of the investigation team.

This case study concentrates on supervised learning models for fraud detection, utilising claims data that is categorised as either 'fraudulent' or 'non-fraudulent'.

Model training

Temporal relevance

It is critical to ensure that training data accurately represents current fraud trends. One effective method is to implement time-based data splits, training the model on historical data while evaluating it on more recent examples. This strategy allows for a better evaluation of the model's ability to generalise in real-world scenarios.

Feature preparation

Features in claims data such as the frequency, claim history the interval between loss and policy inception could have a logical connection to fraud patterns. Subject matter experts should be involved in evaluating the appropriateness of these features as indicators of fraudulent activities.

Imbalance datasets

Instances of fraud could be uncommon, resulting in imbalanced datasets that can potentially bias the model. To identify these challenges and improve model performance, various methods can be employed, including Synthetic Minority Over-sampling Technique (SMOTE), under-sampling and cost-sensitive learning to create data points for AI model training. These approaches help mitigate risks without altering real-world distributions or creating undesirable dependencies, such as repeatedly sampling claims from the same customer.

Handling missing data

It is recommended to consider the patterns of missing data - whether they are random, systematic, or related to the system itself. Employing data imputation methods such as mean substitution or model-based techniques can enhance both the performance and fairness of the fraud detection model

Data Documentation

Recognising the complexities associated with unstructured and semi-structured data provides a better foundation for assessing model risk accurately. The documentation of data transformations and any adjustments applied guarantees reproducibility, transparency, and traceability throughout the data lifecycle.

Model review techniques - Modelling

Model performance metrics

Model reviewers could utilise various performance metrics to assess the fraud detection model's predictive accuracy and effectiveness. These include:

- **Precision, Recall and F1-score** provide insights on the proportion of true positives in the model outcomes,
- **Receiver Operating Characteristic Area Under the Curve (ROC-AUC)** could be used to measures the model's ability to distinguee between fraudulent and non-fraudulent incidents.
- **Calibration Plots,** also called reliability diagrams, are graphical tools that reviewers can used to visually assess the calibration performance of a probabilistic classification model.

All these metrics help model reviewers to determine if the model remains fit for purpose and maintain a good balance between accuracy of fraud detection and the efforts required for investigation.

Fairness and Bias Testing

Fairness is critical when decisions impact financial trust and customer relationships. Disparities can be identified by comparing metrics like F1-score and fairness measures such as Equal Opportunity and Demographic Parity across demographic groups. Visual tools like bar charts and box plots also help reveal differences between groups. Once detected, disparities can be addressed by data re-sampling or adjusting model parameters.

Model explainability

The explainability of predictive AI models can be achieved by selecting inherently interpretable models. For complex models, additional techniques are recommended to enhance transparency for model review. Common approaches include using metrics like SHapley Additive exPlanations (SHAP) values to quantify feature contributions, Local Interpretable model-agnostic explanations (LIME) to analyze individual anomalous predictions, and Partial Dependence Plots (PDPs) to visualize nonlinear feature interactions.

Model drift mitigation

Ongoing monitoring of critical metrics enables the identification of performance degradation, Establishing alert thresholds allows for early detection of potential issues. For example, declines in the Recall, Precision and F1-score may imply model drift. A change in claims pattern or distribution of the input data may trigger the need for model retrain. Changes in claims data can measured by Population Stability Index and Kullback–Leibler (KL) divergence.

The claims data for retraining should be defined based on rolling windows, fixed schedules, or specific event triggers. When retraining the model, it is important to re-evaluate fraud detection thresholds, review essential features, and revisit model assumptions to ensure continued alignment with business objectives.

Use case 2 - Claims summarisation

AI model type: GenAI model

GenAI models can automatically summarise insurance claims by extracting key facts from documents ranging from receipts to medical reports provided. This offers great potential in cost and process time savings in claims handling. However, the claims summarisation model may struggle with complex or rare claims leading to inaccuracies and misrepresentation of key details.

The intrinsic variability in the quality of claim summaries can mask problems like data leakage, prompt injection, hallucinations and biases, potentially resulting in erroneous claims processing and regulatory non-compliance.

Model review techniques

Data leakage control

Preventing data leakage requires a multi-layered approach across the model lifecycle:

- At the data level, classify sensitive information, apply masking or anonymization and enforce strict access controls.
- During model operation, implement safeguards such as prompt filtering and output validation to ensure responses do not expose confidential content.
- For retrieval-augmented generation, secure document indexing, limit retrieval scopes, and maintain audit logs.
- In training, minimize sensitive data use, apply differential privacy and conduct adversarial testing for leakage scenarios.
- Finally, deploy continuous monitoring with Personally Identifiable Information (PII) detection, abuse prevention, and comprehensive logging to maintain compliance and mitigate risks.

Hallucination risk mitigation

Hallucinations occur when an AI system produces content that appears plausible but is factually incorrect or nonsensical. To mitigate this risk the following model review techniques can be applied to detect and address hallucinations:

- **Groundedness Assessment**
  It evaluates whether the information presented in the AI-generated summary can be directly traced and supported by the original source document. A low "groundedness" score suggests potential hallucinations. This assessment verifies that information is not only present but also correctly stated, without distortion or omission.
- **LLM Self-Assessment (LLM-as-a-Judge)**
  A secondary Large Language Model (LLM) is used as a "critic" for the primary claims summarisation model's output. Frameworks like G-Eval utilize an LLM-as-a-judge approach, employing Chain-of-Thought (CoT) prompts to encourage the claims summarisation model to break down the output into intermediate steps. This allows for a more nuanced assessment of correctness, coherence, tonality, and custom Retrieval-Augmented Generation (RAG), leading to more accurate evaluations compared to traditional metrics.

Iterative fine-tuning and retraining

Fine-tuning is typically introduced at a later stage in the model development lifecycle when the requirements extend beyond basic summarisation to include structured outputs, regulatory compliance and high accuracy across diverse claim types.

It adapts a pre-trained claims summarisation model to better handle diverse claim documents. The process is iterative involving validation-based performance checks, hyperparameter adjustments, and model refinements to optimise accuracy and reliability.

To ensure transparency and compliance, all fine-tuning and retraining artifacts such as metrics, configuration changes and comparative results must be retained. These records provide evidence for pre- and post-deployment assessments, confirming that the deployed model delivers the intended summary quality and format.

Human-in-the-loop

While quantitative metrics provide valuable insights, human judgment remains indispensable in identifying hallucinations and biases that automated methods might miss. This involves assigning a person to review and verify the outputs of the claims summarisation model. Human reviewers can assess the accuracy, relevance, coherence, conciseness and readability of the summaries, providing qualitative feedback that is invaluable for detecting subtle hallucinations or biases. For example:

- **Human-Centric Testing** where human testers are involved during the evaluation phase to simulate real-world interventions and identify instances of hallucination or biased outputs.
- **Ethical Review of Data** where human reviewer conducts ethical reviews to ensure that the data used is aligned with societal values and will not lead to unfair or biased decisions.
- **Evaluation metrics** such as ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation - Longest Common Subsequence) and METEOR (Metric for Evaluation of Translation with Explicit Ordering) can be used to quantify the overlap between the AI-generated summary and a human-written reference summary.

Use case 3 - Claims damage assessment

AI model type: Computer vision model

Automating claim assessment can become a compelling opportunity for insurers seeking shortened processing time, improve experience, and enhance decision-making accuracy. A practical workflow involves enabling policyholder to upload images of their damaged vehicles, after which an AI model evaluates the extent of the damage and support claim magnitude estimation.

Traditional data quality control focuses on numerical data or text, allowing clear definitions of completeness, accuracy, and appropriateness. On the other hand, damage assessment models use variable images and require visual inspection for evaluation. Poor image quality directly impacts prediction reliability. The key risk for insurers is that the model may overestimate damage severity, leading to false claims or increased compensation.

Model review techniques

Model Training

Training high-performance vision models from scratch typically requires large domain-specific datasets and extensive compute resources, making it both time-consuming and costly. Consequently, the industry has adopted pre-trained computer vison models – including You Only Look Once (YOLO) , Single Shot MultiBox Detector (SDD), Faster R CNN, and Restnet , EfficientNET or Vision Transformer (ViT) as foundational components. These models have been trained on millions of images and offer robust features extractors that can be adapted to insurance specific tasks.

To tailor these models to claim assessment domain, we leverage transfer learning, a technique that fine-tunes pre-trained architectures using a smaller domain-relevant dataset.

The following techniques can be used enhance the effectiveness of transfer learning strategies:

- **K-fold cross validation** - to measure generalisability and level of model bias
- **Stratified validation** - to ensure that the distribution of target classes in the training and test datasets reflects the original dataset's proportions.
- **Data augmentation** - to expand the size and diversity of dataset. The suitability of the additional data points used should be properly justified.

All of these techniques help to improve model generalisation to real world conditions.

Human-in-the-loop is equally essential to evaluate for marginal cases and challenging scenarios, such as images with low-light conditions or obscured damage, and dynamically monitor and retrain the model as necessary.

Safeguard against overfitting

Techniques such as early stopping and check-pointing further safeguard against overfitting and ensure that the system preserves optimal model weight for deployment.

- **Early stopping** - halts training when model performance no longer improves.
- **Checkpointing** - saves models at specific intervals during the training process. This allows model reviewer to inspect the model performance at various stages and verify the version of model used is optimal and free from overfitting.

Evaluation metrics

Monitoring metric such as validation loss, accuracy and precision recall curves provides insight into model performance during training. Model effectiveness can be validated by assessing the trade-off between precision and recall across thresholds to identify positives while minimising false positives.

Confusion matrix can also be used to explain model accuracy visually. The model reviewer can define and assess the confusion matrices by degree of vehicle damage, showing the false positives and false negatives within each class. This helps model reviewer to identify areas of weakness in the model.

Intersection over Union (IoU) is a metric used in computer vision model. In the case of claims damage assessment, the accuracy of the model can be verified by measuring how well a predicted bounding box of damaged vehicles within an image aligns with the actual object.

Conclusion

The review of AI models requires a bespoke approach depending on the model and use case, extending beyond traditional model validation techniques to address an expanded risk landscape. Given the current maturity, a "human-in-the-loop" remains critical across all use cases.

Insurers should strategically adapt their MRM frameworks, investing significantly in upskilling Line 1 and Line 2 personnel in AI risks and model review techniques, fostering cross-functional collaboration and acquiring new tools to support model review.

To avoid becoming a bottleneck and to ensure confident AI adoption, Line 2 would benefitfrom accelerating enhancing its capability.

Furthermore, it is essential to clearly define performance through established thresholds for model drift and specific metrics, recognising that AI models demand more dynamic and continuous oversight and recalibration than traditional models.

Contact Us

Harvard Lee

Actuarial Director,
Insurance MRM Lead

E: harvard.lee@kpmg.com

Richa Mathur

Senior Manager,
ERS Actuarial Life

E: richa.mathur@kpmg.co.uk

Adedeji Oluwafemi

Data & Analytics Manager
T&D Data Int Ana AI

E: Oluwafemi.Adedeji2@kpmg.co.uk

Shijing Guo

Data Science Senior Manager
T&D Data Int Ana AI

E: shijing.guo@kpmg.co.uk

Stephanie Leung

Manager,
ERS Actuarial Life

E: Stephanie.Leung@kpmg.co.uk

CREATE: CRT165513A | January 2026