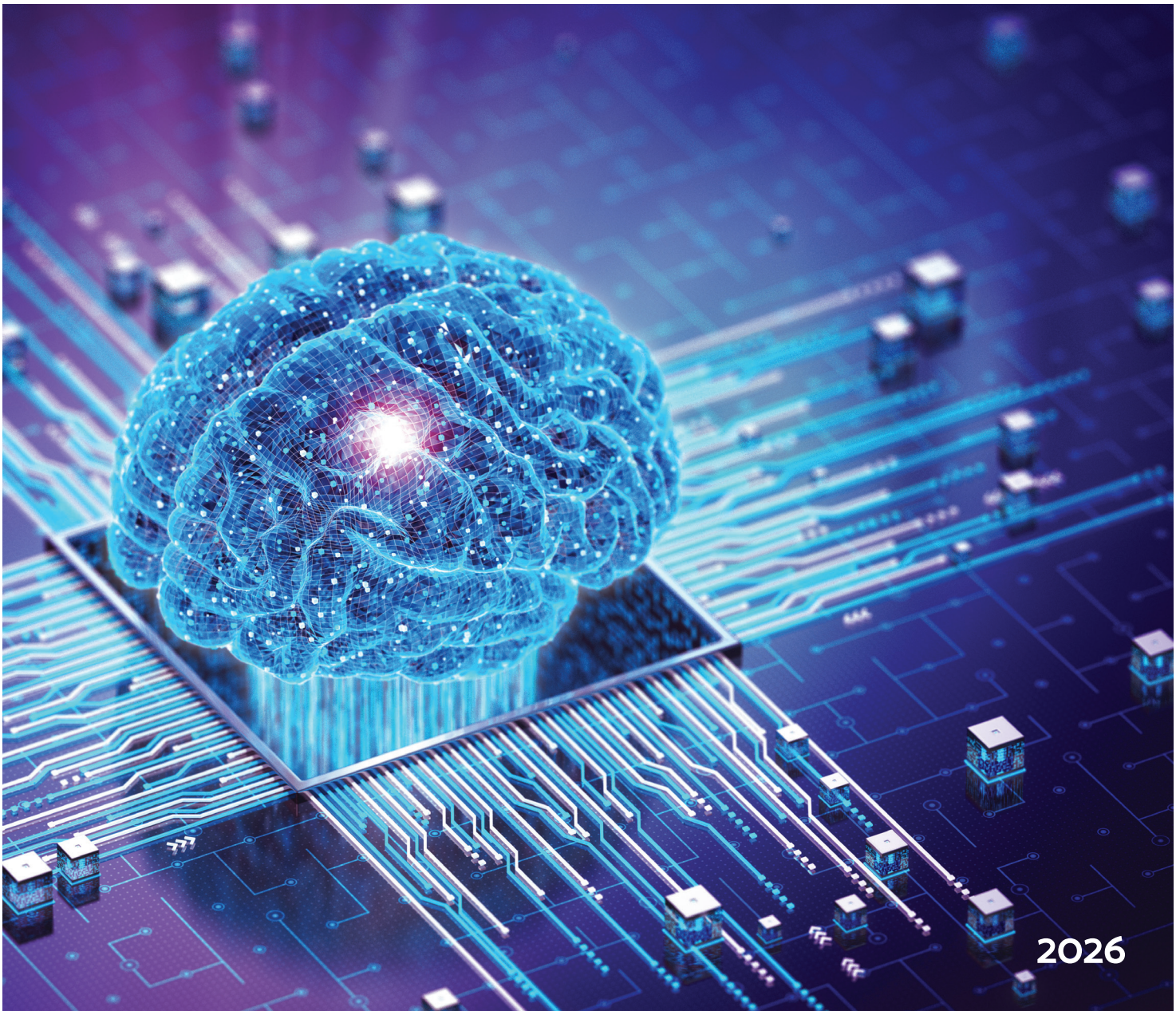
The CAS AI Primer: Practical Guidance for Actuaries

By Shine Wang, Morgan Bugbee, Mario DiCaro, Kris DeFrain, Brooke Engel, DJ Falkson, Bobby Jaegers, Mindy Moss, Christopher Smerald, Xuan You, with contributions from other members of the CAS Artificial Intelligence Working Group



About the Casualty Actuarial Society and the Artificial Intelligence Working Group

The Casualty Actuarial Society (CAS) is a leading international organization for credentialing and professional education. Founded in 1914, the CAS is the world's only actuarial organization focused exclusively on property and casualty risks and serves over 11,000 members worldwide. CAS members are experts in property and casualty insurance, reinsurance, finance, risk management, and enterprise risk management. Professionals educated by the CAS empower business and government to make well-informed strategic, financial and operational decisions.

The CAS Artificial Intelligence Working Group was established to fulfill the CAS mission to "advance the body of knowledge" on a technology that is transforming actuarial practice. Its objective is to encourage the exploration of Artificial Intelligence in actuarial practice through research to help educate members, build knowledge, provide practical insight and establish CAS as thought leaders.

Contents

Why an AI Primer?

Artificial intelligence (AI) is transforming how actuaries work, analyze data, and deliver insights. From automating data processing to developing predictive models and supporting strategic decision-making, AI offers tremendous potential to enhance efficiency, accuracy, and business impact across the insurance value chain.

However, as AI tools – particularly generative AI (GenAI) – become more embedded in actuarial workflows, they also introduce new categories of risk and governance challenges. Understanding these dimensions is critical to ensuring that the adoption of AI is responsible, effective, and compliant.

This AI Primer provides a starting point for actuaries in their AI adoption journey. Specifically, it will:

- Provide a concise overview of AI concepts and applications that are relevant to actuarial work.
- Highlight potential risks.
- Outline best practices for responsible AI use.
- Summarize the key corporate and regulatory considerations that shape AI implementation in actuarial contexts.
- Direct readers to trustworthy learning resources so they may build deeper AI literacy and further develop practical skills.

Throughout this document, the term AI refers primarily to general-purpose large language models (LLMs), such as ChatGPT.

“AI offers tremendous potential to enhance efficiency, accuracy, and business impact.”

Adopting AI in Actuarial Practice

1. Specialize the AI model

In general, LLMs are deployed within companies to assist with low-risk tasks, like writing emails and memos and generating code based on existing ideas. However, most companies are exploring ways to go beyond using AI for administrative or basic tasks. For instance, they may want to explore using AI to clean unstructured data or to develop AI agents to perform routine tasks like triangle development.

One way to investigate where AI could potentially improve business operations is to examine the gap between team or organization goals and the current process used to achieve those goals. For example, perhaps the team needs to categorize claims based on a free-text "loss description" into finer segments for data analysis and modeling to uncover underlying trends that are difficult to spot. AI could be deployed to accelerate the identification of these trends.

Using general LLMs, such as ChatGPT, Llama, or Gemini, without tuning could be problematic because the model would be deriving an answer from the entire universe of data and information but without a focus on actuarial knowledge. Actuaries can tailor a general LLM for actuarial work by using the following methods:

1. **Prompt engineering and context engineering:** Craft effective prompts and context to guide the LLM toward more desired behavior. For example, we can specify in the prompt that the model is an "actuarial pricing expert in an insurance company specializing in auto insurance in California." Such prompts can let the LLM quickly adapt to the specific profession without adding additional cost and effort to adjust the model. However, because it only assigns a role to the model rather than modifying its "knowledge," the output is still unpredictive and harder to apply to business needs.
2. **Retrieval-augmented generation (RAG):** RAG combines an LLM with an external knowledge base that retrieves relevant documents before the model generates a response. We can connect the model to actuarial guidelines, underwriting manuals, claim databases, etc. In addition to the context described in the prompt, RAG provides a company- or profession-specific reference for the LLM to generate better answers without adjusting the model. It is also easy to update the model as the reference material is independent of the model training data. However, it has been reported that the retrieval quality depends on the organization and quality of the information provided (Orofino 2025).
3. **Fine-tuning the LLM:** A general LLM can be retrained by adjusting its internal parameters using domain-specific data. In this process, the model learns directly from actuarial and insurance-related texts – such as reports, filings, and claims documentation – to achieve optimized performance within the actuarial domain. This approach delivers highly specialized behavior and improved accuracy but requires technical expertise, large high-quality datasets, and strict compliance oversight to ensure data security and regulatory compliance.

2. Validate the model

All modeling processes should include result validation. Most LLMs tend to "hallucinate," which is when the model generates false or misleading information and presents it as fact. Model results should be monitored and reviewed by subject matter experts to determine the model's reliability. It is important to understand that an LLM produces results based on a draw from a probabilistic distribution and the results are subject to variability.[1] These models should always be monitored to ensure that they comply with all applicable laws and regulations.

A validation system can consist of one or more validation methods:

- **Human-in-the-loop:** confirm results with human experts.
- **Cross-validation:** compare results across different LLMs.
- **Prompt sensitivity testing:** rephrase prompts to see if the model delivers stable, consistent results.

1 This is unless the "temperature" of the model is set to 1, at which point the model will choose to return the most probable value next. This can lead to nonsensical results.

3. Avoid AI pitfalls

GenAI is a powerful assistant, but it is not a substitute for critical thinking. One of the biggest risks with using AI lies in an overreliance on AI-generated outputs without adequate scrutiny. Actuaries feeling the pressure of deadlines or who assume that "AI knows better" may be tempted to accept AI responses at face value rather than applying their own professional judgment and analytical reasoning. Relying on these outputs without critical evaluation can lead to flawed assumptions, misinterpretation of trends, or – in the most extreme cases – even regulatory noncompliance.

To mitigate these risks and use AI responsibly, actuaries should adopt a mindset of continuous validation. This involves establishing validation loops in which AI outputs are tested, reviewed, and refined over time. Comparing results against traditional models, conducting sensitivity analyses, and peer-reviewing AI-generated insights are effective ways to ensure outputs are technically sound and contextually appropriate.

As the AI industry and technologies rapidly evolve, there is no well-defined roadmap to success. Instead, success will come to those who experiment, collaborate, and keep business objectives top of mind.

Requirements for Corporate AI Implementation

The adoption of AI needs to suit the corporation's needs and strategic blueprint. Potential use cases and risks in both frontline and backend operations need to be separately identified. On the frontline, AI can enhance efficiency by automating manual processes, extracting insights from unstructured data (such as adjuster notes or legal documents), and improving customer experience through personalization and real-time support. However, mistakes on the frontline could directly impact customer experience. On the backend, AI can streamline data collection, improve algorithmic performance, automate reporting, and provide faster, more insightful interpretations to support decision-making. Mistakes on the backend could distort data used for decision-making, ultimately impairing company performance.

A robust governance framework that identifies and mitigates distinct model risks across various AI applications is imperative for successful AI development and widespread adoption. It should take into consideration the unique risks and overall strategic goals of each individual corporation. Key elements of this framework include, but are not limited to, regulatory compliance, mitigation of model bias, and clear model accountability. Additionally, documentation, training, and standardized processes must be prioritized to ensure consistency and auditability throughout the model development and deployment lifecycle. Teams should be equipped with a solid understanding of AI risks and best practices, particularly if they are expected to use or interpret AI outputs.

High-quality, representative data is foundational to any AI initiative. This requires strong data readiness practices, including metadata management, data lineage tracking, and enterprise-wide data quality frameworks. Equally important are data security and privacy. Sensitive or personal data should be de-identified where necessary, and appropriate encryption and access controls must be in place. Organizations have a key responsibility to improve overall data literacy among their employees and foster responsible data use, which

"A robust governance framework that identifies and mitigates distinct model risks across various AI applications is imperative for successful AI development."

is especially important when AI models are embedded in decision-making processes.

Another critical decision involves model ownership – whether to build AI models in-house or use those provided by vendors. Traditionally, in-house models are highly customizable, transparent, and explainable, while vendor models are less flexible but easier to maintain. However, this distinction becomes blurred with GenAI. Most organizations do not have the infrastructure or data to build GenAI models from scratch and instead rely on third-party foundation models that they fine-tune or enhance using techniques like RAG. Due to the complexity and high-energy consumption of GenAI, corporations need to rely on cloud computing to fully utilize these models. Privacy and compliance – both internal and external – can be major concerns in such applications, given the amount of sensitive insurance information that companies maintain. In addition, these models often function as "black boxes," even to those who developed them. Mistakes are harder to detect compared with other machine learning models and may have more severe consequences. Therefore, well-defined evaluation criteria that align with the corporation's risk appetite must be applied to assess both the data used in the model and the model's outputs.

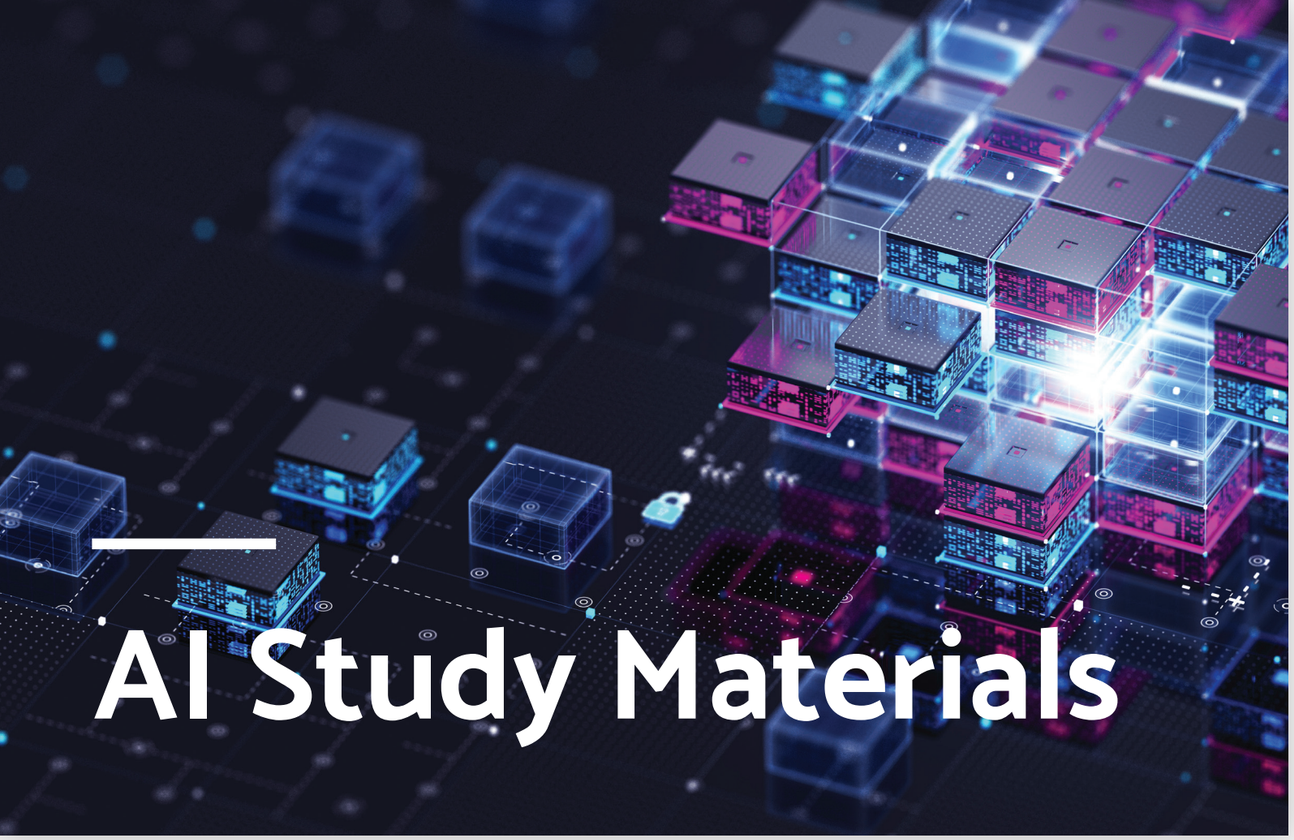
AI models are not static assets. They degrade over time due to changes in data patterns, customer behavior, or external conditions. Companies must plan for ongoing monitoring and model maintenance. This includes:

- implementing dashboards to track performance metrics, such as accuracy, drift, and fairness;
- setting retraining schedules or triggers to update the models as needed; and
- developing fallback protocols in case the models fail or become unreliable.

Without these safeguards, AI systems can quickly lose effectiveness and pose reputational or regulatory risks.

To make things more complicated, the above steps do not work in isolation. AI models are used to improve data quality and augment current company databases. These additional data sources then feed into other models used for decision-making, which could in turn affect future data requirements. Such an iterative process could expose companies to compounding risks from AI models. As a result, company-wide standards and industry-wide oversight are critical for a responsible adoption of AI. Within the corporation, interdisciplinary collaboration is essential. Risk management must be involved in assessing and overseeing model risk; legal and compliance teams must guide the ethical and lawful use of AI; IT must provide the infrastructure and security layers; data teams must manage inputs and pipelines; and business units must define use cases and constraints. Internal audit also plays a critical role in ensuring proper governance and oversight. By involving all these stakeholders early and often, companies can build AI solutions that are not only technically sound but also trustworthy, compliant, and aligned with strategic goals.

Finally, organizations should factor in the environmental and sustainability impacts of AI. Large models, particularly GenAI, require substantial computing resources, which translate into a significant carbon footprint. Choosing efficient architectures, using cloud providers with sustainability commitments, and balancing model complexity with business value are all essential components of responsible AI implementation.



AI Study Materials

Expand your AI topic knowledge and experience, from foundational through advanced, with these resources.

For those who may feel overwhelmed by the AI landscape, the CAS Institute (iCAS) AI Fast Track bootcamp demystifies AI and leverages your existing skills to help you explore transformative applications in actuarial science, gain practical insights into key AI techniques, and discover how to integrate them into your workflow to enhance efficiency and innovation. An on-demand version of the bootcamp is now available.

Once you have learned the basics to practice in a safe and secure environment, learning by doing is one of the best ways to quickly upskill. Ask the GenAI tools how they can assist you and spend 15 minutes a day exploring the possibilities. This is an effective way to expand your knowledge even if you don't have a lot of time to devote to formal learning.

The CAS and iCAS offer a wide range of AI-related content and education for actuaries of all experience levels. In the September/October 2025 issue of *Actuarial Review*, Dan Jackman shared an AI compendium, organized by topic (Jackman 2025). This helpful resource lists articles, podcast episodes, webinars, event sessions, courses, and bundles covering many aspects of AI in an actuarial science and property and casualty insurance context.

1. Introduction to AI in insurance

Resources for those newer to AI or who are seeking foundational content.

Title	Type	URL
A Focus on Research and Volunteers	Article	https://ar.casact.org/a-focus-on-research-and-volunteers/
Actuarial Workflows with AI	Course	https://www.pathlms.com/cas/courses/90048/video_presentations/332504
Agentic AI: Your New Actuarial Coworker	Article	https://ar.casact.org/agentic-ai-your-new-actuarial-coworker/
AI Compendium – iCAS	CAS resource	https://thecasinstitute.org/professional-education/ai-compendium/
Almost Nowhere Episode 1: Joshua Meyers	Podcast	https://open.spotify.com/episode/3Qy8mGBvIK45IHG6cIRwdT?si=5kqmyXtHQGeqiRxO_-WjVQ
How Actuarial Science Can Benefit from AI ... and Vice Versa	Session	https://www.pathlms.com/cas/events/12068/event_sections/17771/video_presentations/356916
Intersection of Actuarial Science and Artificial Intelligence	Webinar	https://www.pathlms.com/cas/courses/74642

2. AI in pricing, reserving, and underwriting

Materials focused on practical modeling and core property and casualty functions.

Title	Type	URL
AI Insurance: Managing and Underwriting Enterprise AI Risks	Session	https://www.pathlms.com/cas/events/12068/event_sections/17770/video_presentations/356909
Artificial Intelligence Gone Nuclear	Article	https://ar.casact.org/artificial-intelligence-gone-nuclear/
Spring 2025 AI Bundle	Bundle	https://www.pathlms.com/cas/product_bundles/15497
Rapidly Evolving Technology and Its Implications for the Reserving Process	Article	https://ar.casact.org/rapidly-evolving-technology-and-its-implications-for-the-reserving-process/
Reserve in Machine Learning, 2025 iCAS Data Science & Analytics Forum	Session	https://www.pathlms.com/cas/courses/104478/video_presentations/348393
Risk Evaluation for a Cloud-Based AI Model	Session	https://www.pathlms.com/cas/events/12068/event_sections/17770/video_presentations/356912
The Use of AI in Insurance	Session	https://www.pathlms.com/cas/events/11818/event_sections/17371/video_presentations/347814

3. Risk management, enterprise risk management, and strategy

AI resources for enterprise risk and capital decisions.

Title	Type	URL
AI-Empowered Actuaries: An Introduction to AI Agents	Session	https://www.pathlms.com/cas/events/12068/event_sections/17770/video_presentations/356950
Application of AI and Machine Learning in (Re)Insurance	Session	https://www.pathlms.com/cas/events/12221/event_sections/17909/video_presentations/361288
ERM: Using AI in Scenario and Stress Testing for Optimizing Insurance Strategy (Part 1)	Session	https://www.pathlms.com/cas/events/10243/event_sections/16507/video_presentations/328264
From AI to Climate Risk: Updates from the Recent IAA Meeting in Tallinn, Estonia	Article	https://ar.casact.org/from-ai-to-climate-risk-updates-from-the-recent-iaa-meeting-in-tallinn-estonia/

4. Ethics, regulation, and responsible AI use

Key study materials for actuaries working with governance, compliance, and policy.

Title	Type	URL
AI Generates Single Point of Failure Rethink	Article	https://ar.casact.org/ai-generates-single-point-of-failure-rethink/
AI Regulation in Insurance: A Road to Unintended Consequences	Article	https://ar.casact.org/ai-regulation-in-insurance-a-road-to-unintended-consequences/
AI: A Multi-faceted Cyber Threat	Session	https://www.pathlms.com/cas/events/12068/event_sections/17771/video_presentations/356915
Almost Nowhere Episode 6. Jim Guszcza	Podcast	https://open.spotify.com/episode/1FKXH9kXPkuhVkcqvig8Li?si=xpBWktRoRgeuKw2kRORqsg
Bridging Data Divides: AI as a New Paradigm for Unstructured Data	Session	https://www.pathlms.com/cas/events/11818/event_sections/17370/video_presentations/347807
GenAI Related Litigation Brings Fair Use into Focus	Article	https://ar.casact.org/genai-related-litigation-brings-fair-use-into-focus/

5. The actuary of the future and transformation of roles

Conversations about how AI is reshaping actuarial identity and skillsets.

Title	Type	URL
Four Futures for Actuaries in the Wake of AI	Article	https://ar.casact.org/four-futures-for-actuaries-in-the-wake-of-ai/
Almost Nowhere Episode 7. Frank Chang	Podcast	https://open.spotify.com/episode/20P2bA3SiW4dXI48biTwvS?si=zaOtzHc0RIW1SRMIgk4HAQ
Almost Nowhere Episode 8. Charlie Stone & Brian Fannin	Podcast	https://open.spotify.com/episode/5mATz1FsAGoyDMZHc2Ao2y?si=kg70cmH1ROOPAnxbPMWcGw
Tech for Pros: An Overview of Modern Ops	Session	https://www.pathlms.com/cas/courses/104478/video_presentations/348395

6. Infrastructure, operations, and the AI ecosystem

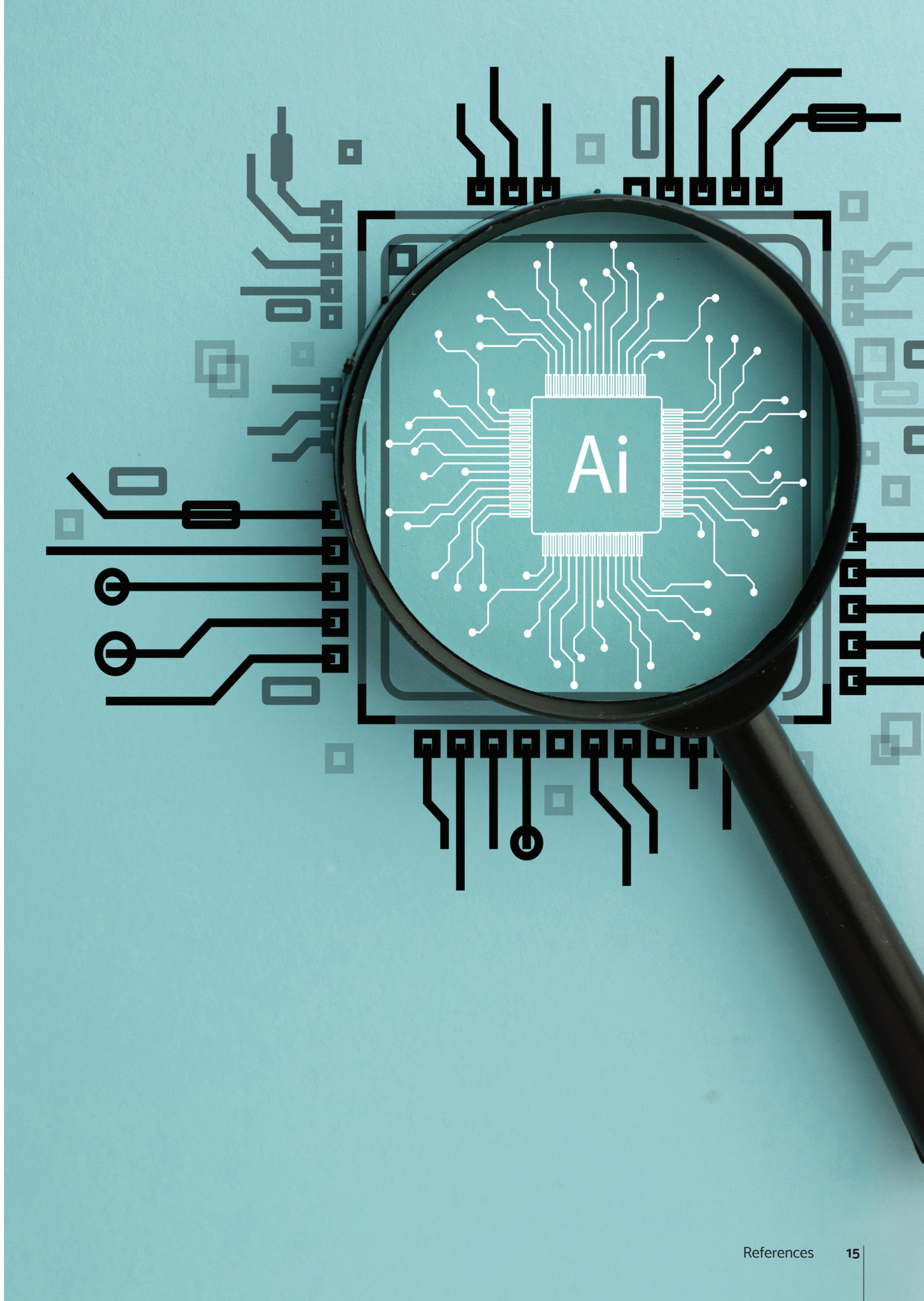
Resources for those interested in technical stacks, systems, and deployment.

Title	Type	URL
AI Fast Track On-Demand Course	Course	https://www.pathlms.com/cas/courses/113452
Almost Nowhere Episode 2. Sergey Filimonov	Podcast	https://open.spotify.com/episode/3FVgt2rsxnkrpMd7nm7ijd
Bridging Data Divides: AI as a New Paradigm for Unstructured Data	Session	https://www.pathlms.com/cas/events/11818/event_sections/17370/video_presentations/347807

References

Jackman, Dan. 2025. "The AI Moment in Insurance: How Actuaries Are Grappling With the Future." *Actuarial Review*, September 2025. https://ar.casact.org/the-ai-moment-in-insurance-how-actuaries-are-grappling-with-the-future/

Orofino, Diego. 2025. "RAG vs. Fine-Tuning: Which One Should You Use for Your AI Model?" *Agentic AI Minds*, March 2025. https://medium.com/agentic-minds/rag-vs-fine-tuning-which-one-should-you-use-for-your-ai-model-8532a8552fd3



Appendix

1. Key terms and definitions in GenAI

Actuaries new to GenAI may encounter technical terms. This glossary is tailored for those newer to LLMs.

Term	Definition
agentic capabilities	The ability of a model to plan multistep solutions, use tools such as code or search, and act more like a junior analyst.
benchmarks	Standard tests to compare models: • MMLU: Exam-style questions across disciplines. • SWE-Bench: Software and coding tasks. • AIME: Math problem-solving. Benchmarks indicate strengths but do not guarantee accuracy on actuarial work.
closed-weight versus open-weight	Closed-weight models (e.g., GPT-5, Claude, Gemini) are accessed only through the developer's systems. Open-weight models (e.g., Llama 4) can be run privately on one's own infrastructure.
context window	The number of tokens the model can consider at once (input + output). A larger window means you can provide longer documents, such as rate filings or full claim histories.
mixture of experts (MoE)	A model design in which only part of the model activates for each prompt, making it more efficient.
multimodal	A model that can handle inputs beyond text (e.g., images, audio files, spreadsheets).
time to first token (TTFT)	The amount of time it takes for the model to start responding. Larger or more reasoning-oriented models take longer.
token	A unit of text (roughly 3/4 of a word). • Input tokens are what you send to the model. • Output tokens are what the model produces.
tokens per second (t/s)	The speed of output generation, measured in seconds.

2. Comparison of leading GenAI models

Model	Description	Key Metrics*	Insurance Use Cases	Helpful Skills	How to Use and Caveats
GPT-5 (OpenAI)	Flagship model with advanced reasoning and "thinking mode." Excels at summarization, coding, and communication.	**Context:** Up to 400k **Price:** ~$1.25/M input, $10/M output **Speed:** ~134 t/s **Benchmarks:** Top tier on SWE-Bench and AIME	Drafting rate filing language; summarizing bureau circulars; creating management decks; coding automation tasks.	Clear prompts; ability to check generated Python/R code; validation of regulatory language.	Accessed via ChatGPT or API. Reasoning mode improves accuracy but increases cost and latency.
Grok 4 (xAI)	Speed-focused model with built-in live web search.	**Context:** 256k **Price:** ~$3/M input, $15/M output **Speed:** ~50–56 t/s **Benchmarks:** Mid-range SWE-Bench	Monitoring competitor filings; summarizing customer complaints; scanning catastrophe news for underwriting impact.	Effective at validating sources and interpreting web content.	Accessed via API. Web search provides current data but must be verified.
Gemini 2.5 Pro (Google)	Long-context model (up to 1M tokens) integrated with Google Workspace (Docs, Sheets, Gmail).	**Context:** Up to 1M **Price:** ~$1.25–$2.50/M input, $10–15/M output **Benchmarks:** Strong on long-context and reasoning; strong in document synthesis	Synthesizing submissions to the System for Electronic Rates & Forms Filing (SERFF); comparing assumptions across states; analyzing long claims histories; summarizing Special Investigation Unit (SIU) investigations.	Organizing large datasets; effective working with Docs/ Sheets integration.	Available via Google Cloud (Vertex AI) or directly in Workspace apps. High token use increases cost.
Claude 4 Opus (Anthropic)	Premium reasoning-oriented model. Known for following instructions carefully, delivering structured outputs, and providing coding support.	**Context:** Up to 1M **Price:** ~$15/M input, $75/M output **Benchmarks:** High on reasoning and SWE-Bench	Drafting validation reports; extracting information from policy forms; coding automation; challenging actuarial assumptions.	Providing schemas and examples; effective with structured outputs (tables, JSON).	Accessed via Anthropic or AWS Bedrock. Costs rise quickly with long documents.
Llama 4 Maverick (Meta)	Open-weight, multimodal model, designed for enterprise self-hosting. Offers privacy and customization.	**Context:** ~1M **Price:** Hosting dependent (often <$1/M) **Benchmarks:** Solid coding and reasoning for its size	Privacy-sensitive extraction (policy forms, personally identifiable information [PII]); batch processing of claims; in-house actuarial Q&A tuned on company documentation.	Basic infrastructure and fine-tuning skills; retrieval augmentation.	Must be hosted internally or through a vendor. Performance depends on hardware and fine-tuning quality.

*Metrics vary by provider and workload. Values shown are representative as of August 2025.

3. Matching model to task (quick guide)

Coding and automation (Python, R, Excel macros): GPT-5, Claude 4 Opus

Long filings and regulatory synthesis: Gemini 2.5 Pro, Claude 4 Opus

Market and competitor monitoring: Grok 4

On-premises privacy-sensitive work: Llama 4 Maverick

General communication and documentation: GPT-5, Claude 4 Opus

4. Prompting and practical tips

GPT-5 (OpenAI): GPT-5 is a capable coding assistant for actuaries, converting Excel or VBA routines into Python or R, debugging reserving scripts, or generating unit tests for pricing models. It is also adept at producing structured tables or JSON outputs for filings and can chain tasks such as summarizing a filing, anticipating regulator objections, and drafting responses. It is best used for both code and communication support, with "thinking mode" reserved for more complex reasoning tasks.

Grok 4 (xAI): Grok 4 delivers strong performance on prompts requiring current information, such as competitor filings or regulatory updates, and can generate supporting scripts to parse and analyze data. Always request sources from Grok 4 output, as its live-search ability is powerful but requires validation. It is particularly useful for scanning market activity, complaint trends, or catastrophe developments that actuaries can integrate into their analyses.

Gemini 2.5 Pro (Google): Gemini 2.5 Pro is well suited for actuaries who work heavily in Excel and Google Sheets, bridging spreadsheet logic with Python coding and producing outputs directly back into Sheets. Its very large context window enables entire filings or claim histories to be processed at once. Workspace integration allows for seamless drafting in Docs or Sheets, making it approachable for actuaries who prefer using familiar tools.

Claude 4 Opus (Anthropic): Claude 4 Opus provides careful, transparent outputs, making it valuable for reviewing actuarial code, drafting validation reports, or producing structured extraction routines from policy forms. It can generate "challenge memos" that combine coding suggestions with actuarial reasoning, supporting governance and model risk processes.

Llama 4 Maverick (Meta): Llama 4 Maverick is a cost-efficient, open-weight option that can be run internally for privacy-sensitive work. When fine-tuned on actuarial libraries, it can function as an internal copilot for rating scripts or claims models or power secure batch jobs. While not as strong on general reasoning as GPT-5 or Claude, it is ideal in situations where control and customization are priorities.

5. Practical realities

Benchmarks ≠ business results. Test models on actuarial tasks, such as assumption reviews or filing drafts, before adoption.

Reasoning has a cost. "Thinking modes" improve reliability but slow response and increase expense. Use them for complex problems.

Open models are not "free." They reduce API (application programming interface) costs but shift responsibility for infrastructure and maintenance.

Leaderboards are imperfect. GenAI models show relative strengths but not accuracy on insurer-specific data.

Expertise. Insight.
Solutions.®

Casualty Actuarial Society
4350 North Fairfax Drive, Suite 250
Arlington, Virginia 22203
casact.org