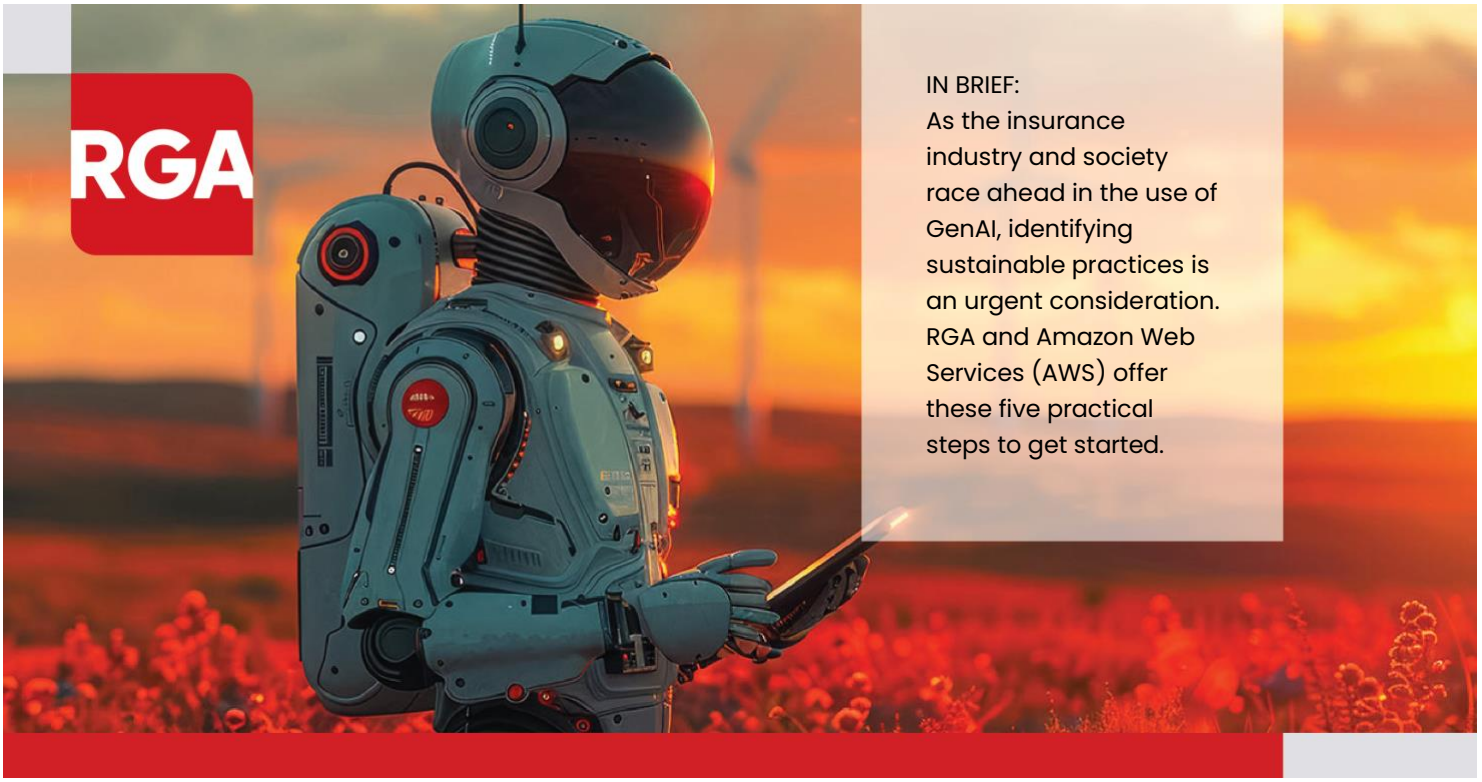

5 Ways to Reduce GenAI's Carbon Footprint

By Jeff Heaton, Armana Saffie, Tony Howell

KEY TAKEAWAYS

- GenAI's role in insurance will only continue to grow, and insurers must develop sustainable practices today to prepare for the insurance industry of tomorrow.
- Five ways to reduce AI's carbon footprint include AI infrastructure optimization, data efficiency, model optimization, user training, and governance.
- A holistic, lifecycle-based approach to AI that integrates sustainability into every application can help insurers align operational efficiency with sustainability goals.

Generative artificial intelligence (GenAI) offers numerous advantages, enhancing productivity and streamlining operations. However, as GenAI adoption grows, addressing its environmental impact is crucial.

This article, a collaboration between RGA and AWS, highlights five key areas to minimize GenAI's carbon footprint. From individual users to organizational leaders, everyone has a role in ensuring the sustainable use of this powerful technology.

1. AI infrastructure optimization

One of the most direct steps to reduce the environmental impact of AI is selecting energy-efficient infrastructure. Moving GenAI workloads to a hyperscale cloud provider is an increasingly popular choice. AWS, for example, operates data centers designed for high efficiency, with advanced cooling systems, custom silicon (e.g., Graviton processors and AI accelerators such as Trainium and Inferentia), and architectural innovations that minimize energy loss. Consolidating resources at scale spreads operational costs across multiple users, providing economic and environmental benefits.

Many cloud providers are actively pursuing net-zero emissions. AWS achieved its 100% renewable energy goal in 2023, seven years ahead of its original target of 2030, as part of Amazon's broader Climate Pledge to reach net-zero carbon by 2040.

> From an insurer's standpoint, choosing a region or data center with a high percentage of renewable energy can dramatically lower the carbon footprint of AI workloads. And because cloud-based services are elastic, insurers can scale up or down based on demand rather than running idle, power-consuming servers on premises.

In addition, managed services such as Amazon Bedrock can streamline large language model (LLM) deployment and optimization. Because these services handle infrastructure adjustments, insurers no longer need to fine-tune hardware themselves, avoiding common pitfalls such as over-provisioning. This not only boosts performance and reliability but also curbs energy consumption.

2. Data efficiency

AI performance depends on quality data, yet more data does not necessarily lead to better results. For insurers, focusing on data curation, cleaning, and annotation often proves more effective than simply feeding models massive amounts of unrefined information. High-quality data helps models learn patterns faster, requiring fewer training cycles and, consequently, less energy. Because insurers deal with extensive customer data, from underwriting to claims, the payoff in resource savings and accuracy can be substantial if data is carefully prepared from the outset.

Partner with RGA to responsibly combine advanced technology like artificial intelligence, data, and expertise to deliver value to your customers.

Learn more →

3. Technical considerations/model optimization

Insurers use AI for tasks ranging from assessing claims to detecting fraud. Though large, general-purpose AI models are flexible, they can be excessively energy hungry. By contrast, smaller or more task-specific models might deliver comparable accuracy while consuming far less energy. Whether leveraging models fine-tuned for reliability or using distilled versions for real-time risk scoring, running leaner algorithms can help organizations meet both performance targets and sustainability goals.

Another practical measure is managing "chat memory." For example, if a model powers a policyholder support chatbot, a large context window means the system processes each new inquiry against a long history of previous messages. Not only does this require more tokens to be processed, but it also increases inference time and energy consumption. Limiting context windows, or starting fresh for each new task, reduces this overhead and can improve response quality.

For non-real-time tasks, such as summarizing large batches of documentation, it is often more efficient to process data in bulk without any conversational memory. Retaining a chat history for each piece of data would quickly fill up context buffers and undermine accuracy. Similarly, prompt caching, though still maturing, shows promise in reducing redundant computation.

> By reusing computed states for common prompts, insurers can lower the energy and time needed to handle repeated inquiries, especially relevant when multiple agents or applications rely on similar prompts.

Quantization is another valuable tool. By reducing the numerical precision of a model's weights from 32-bit to 8-bit integers, quantization shrinks the overall model size and cuts down the time and energy required to run each query. For tasks where absolute precision is not essential, this trade-off can yield large energy savings with minimal accuracy loss.

4. User training

User behavior significantly affects the energy footprint of AI systems. In claims processing chatbots, for example, employees and policyholders who frequently switch tasks in a single session can inadvertently balloon the context window. Training users to open a fresh chat session for each distinct inquiry keeps token counts lean and inference times short. This translates to important resource savings for cost-conscious insurers and often more precise, relevant model outputs.

5. Monitoring and governance

Effective AI governance starts with resource-use transparency. Tracking token consumption, setting departmental budgets, and identifying usage "spikes" allow insurers to detect inefficiencies and adjust their strategies accordingly. Tools from providers such as AWS help assign costs to specific teams or lines of business, providing valuable visibility into who is using AI, how much it costs, and whether there are ways to optimize spending.

Additionally, measuring carbon emissions associated with AI workloads is important for broader ESG reporting. Services such as the AWS Customer Carbon Footprint Tool can reveal how infrastructure choices translate into actual environmental impact, giving insurers the information needed to make evidence-based sustainability improvements.

Beyond the model: Holistic efficiency

As GenAI gains traction in insurance, sustainability considerations must be integral at every stage. By adopting energy-efficient cloud services, refining data pipelines, right-sizing models, and training teams on responsible AI practices, insurers can deliver cutting-edge solutions without disproportionately increasing energy consumption. Equally important is a strong governance framework to monitor usage, costs, and carbon emissions, ensuring organizations fulfill their environmental commitments while safeguarding customer data.

A holistic, lifecycle-based approach to AI helps insurers align operational efficiency with sustainability goals. From initial data collection and model selection to the retirement or replacement of outdated systems, each phase offers an opportunity to reduce waste, minimize emissions, and meet rising regulatory and stakeholder expectations. With thoughtful planning and execution, GenAI can be a powerful catalyst for innovation while upholding the insurance industry's commitment to a more sustainable future.

AUTHOR

Jeff Heaton
Vice President, AI Innovation

AUTHOR

Armana Saffie
Assistant Vice President,
Global Accounts

AUTHOR

Tony Howell
Senior Solutions Architect,
Amazon Web Services

Want to learn more about RGA's AI expertise and how our insights can benefit your organization?
Let's talk.