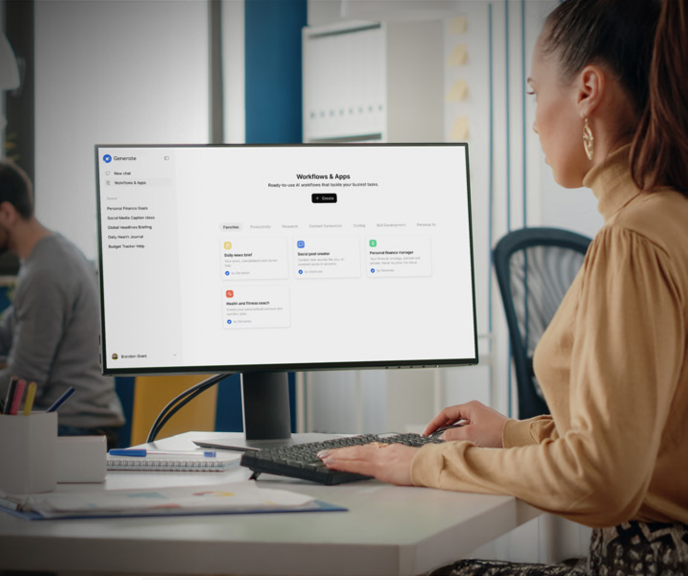
ITERATE.AI BUILDS PRIVATE AI ON AMD RYZEN™ AI PRO PROCESSORS

CASE STUDY

Iterate.ai taps AMD Ryzen™ AI PRO processors for 32B private LLMs with 32k context window models at ~60-80 tokens/sec, cutting cloud costs and risks



Iterate.ai calls itself "the private AI company." As enterprises adopt generative AI to boost productivity, Iterate.ai focuses on sophisticated workflows that run securely within organizational boundaries. Its Generate application enables large organizations to build personalized AI assistants and agentic workflows that efficiently analyze complex data scattered across critical business systems such as Jira, Slack, and Salesforce.

> **"AMD is at the right spot to enable the private AI revolution, making them an ideal partner for Iterate.ai."**
>
> Brian Sathianathan, Co-founder and CTO, Iterate.ai

To deliver powerful local AI functionality while helping prevent data exposure, Iterate.ai collaborated closely with AMD to optimize Generate exclusively for AMD Ryzen™ AI PRO processors. This collaboration helps Generate use the AMD Ryzen AI PRO processor's CPU, integrated GPU, and dedicated NPU in concert. By anchoring the solution to the user's device, Iterate.ai provides 100% data retention and control, while converting variable cloud token expenses into a predictable, fixed cost for enterprises.

SECURING DATA AND CONTROLLING TOKEN COSTS

IT decision makers value both performance and security. The shift to AI intensified security concerns as organizations struggle to manage the risk of sensitive, proprietary information being exposed to public large language models (LLMs).

> **"Iterate.ai runs the AI models right next to the application on your local device so that no data is pushed to the Internet."**
>
> Karanbir Singh, Special Project Engineer at Iterate.ai

"The big pain point that Generate addresses is safety," said Iterate.ai Special Project Engineer Karanbir Singh. "When companies use a public LLM, they don't have control over the data that is retained. Iterate.ai runs the AI models right next to the application on your local device so that no data is pushed to the Internet."

INDUSTRY

Software

CHALLENGES

Enterprises need to use generative AI without exposing sensitive data to public LLMs, while controlling fast-rising token costs and supporting users while traveling or offline.

SOLUTION

Iterate.ai optimized its Generate AI platform for PCs with AMD Ryzen™ AI PRO processors and used AMD Lemonade Server to speed local LLM development and reduce engineering needs.

RESULTS

Generate runs 32B models, delivering 32K-context private models at approximately 60-80 tokens/sec on AMD Ryzen™ AI PRO processor-based PCs, turning volatile cloud token spend into fixed CapEx and keeping data on the device.

AMD TECHNOLOGY AT A GLANCE

AMD Ryzen™ AI PRO processors
AMD Lemonade Server

Compounding the challenge is cost. Iterate.ai Co-founder and CTO Brian Sathianathan says, "Recent industry studies indicate that there has been as much as a 100X increase in token usage in just the past year." Cloud-based LLM inference is charged per token, creating high variability and cost uncertainty for enterprise budgets. Iterate.ai wanted to empower enterprises to move their AI workflows away from this usage-based model and establish full data sovereignty. The solution also needs to work offline so users can continue using AI even when traveling or experiencing connectivity issues.

PRIVATE AI ADVANCED BY AMD POINTS THE WAY

Iterate.ai chose AMD for its comprehensive technology stack, which provides the scale and flexibility needed for its enterprise product. "What's compelling about AMD is that they have a laptop offering with potent processors that include CPU, GPU, and NPU capabilities. At the same time, AMD also offers powerful server-side solutions like the AMD Instinct™ MI300 Series GPUs," says Sathianathan. "With AMD, there's an end-to-end solution from the client side all the way to the server. AMD is at the right spot to enable the private AI revolution, making them an ideal partner for Iterate.ai."

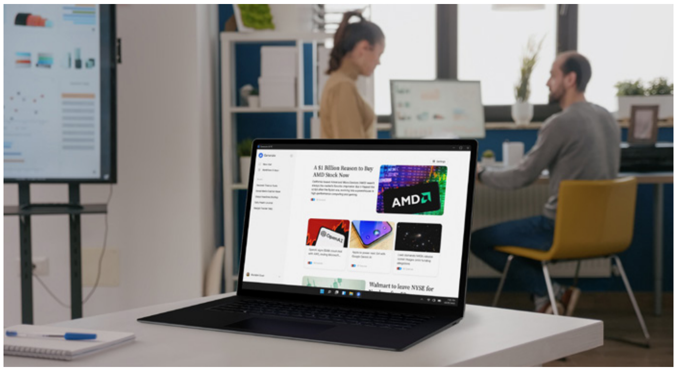
> **"With AMD, there's an end-to-end solution from the client side all the way to the server."**
> Brian Sathianathan, Co-founder and CTO, Iterate.ai

ACCELERATING DEVELOPMENT WITH AMD LEMONADE SERVER

Iterate.ai integrated AMD Lemonade Server, an open-source tool that simplifies development for local LLMs. "Lemonade Server exposes everything as OpenAI compatible, the default industry standard. The work AMD put into Lemonade Server makes integration between our application and the LLM much easier," said Sathianathan.

The streamlined approach accelerated Iterate.ai's deployment timeline and reduced resource requirements. Sathianathan says, "It would have taken us at least twice the number of engineers had we had to do all the work ourselves. The combination of AMD software and hardware is very stable, stellar in fact, and allowed us to manage development with far fewer resources than typically required on the LLM side."

> **"The combination of AMD software and hardware is very stable, stellar in fact, and allowed us to manage development with far fewer resources."**
> Brian Sathianathan, Co-founder and CTO, Iterate.ai



Generate uses the CPU, GPU, and NPU of AMD Ryzen AI PRO processors in parallel to handle everything from document summaries to complex chart analysis.

OPTIMIZING HARDWARE UTILIZATION

The high-performance architecture of the AMD Ryzen AI PRO processor allowed Iterate.ai to design a local AI application with greater capacity than previously possible. Generate uses all three components of the AMD Ryzen AI PRO processor to achieve optimal performance: the CPU, GPU, and NPU. This parallel architecture is essential because Generate handles use cases from simple document summarization to highly complex analysis of charts and graphs.

> **"The AMD Ryzen AI NPU is super-optimized to run the LLM for long periods, which is perfect for very complex tasks."**
> Brian Sathianathan, Co-founder and CTO, Iterate.ai

The coordinated hardware acceleration is vital for power efficiency and battery life. Sathianathan highlighted the benefits of leveraging the specialized NPU, saying, "The AMD Ryzen AI NPU is super-optimized to run the LLM for long periods, which is perfect for very complex tasks. On a PC you live in a memory- and resource-constrained environment. We take advantage of the different power profiles of NPU, GPU, and CPU. Being able to use all of the AMD Ryzen AI PRO processor's silicon resources in parallel is ideal. That approach also helps avoid draining the user's laptop battery during heavy LLM processing."

ACHIEVING SCALE, FIXED COSTS, AND DATA SOVEREIGNTY

Sathianathan notes that deploying Iterate.ai on the AMD platform is essential to their ability to meet the demands of the modern enterprise. "AMD helps make it possible to leverage local AI with security, privacy, and cost benefits, creating certainty for businesses and organizations," he says.

At the same time, such certainty is built on measurable, superior performance. Iterate.ai testing confirmed the AMD platform enabled exceptional model scaling. “In industry standards, 14 billion parameter models with a 16K context window are good benchmark models for an AI PC,” said Singh. “With AMD Ryzen AI processors, we can leverage models that are 32 billion parameter models, almost double the size, with a 32K context window.”

This scale advantage allows Iterate.ai to handle more demanding enterprise workloads locally on the PC. Testing metrics validate the high throughput of the proprietary Iterate.ai model running on the system: its Interplay RAG model achieved an average throughput of ~60-80 tokens/sec, with a peak of 92.89 tokens/sec during a Document Search query.

“Running Generate locally means it’s 100% fixed cost,” says Sathianathan. “You move a lot of variability costs because it’s a CapEx, one-time purchase, that you basically amortize over the life of the PC.”

> **“I think AMD is doing great in their vision for the future, and we’re excited to be their partner.”**
>
> Karanbir Singh, Special Project Engineer at Iterate.ai

By keeping the entire LLM’s memory on the PC, Generate supports data sovereignty and helps organizations align with compliance standards such as HIPAA, GDPR, and SOC 2. This privacy-first approach is reinforced by AMD PRO Security, which provides multi-layered defenses such as AMD Memory Guard. AMD Memory Guard encrypts system memory in real time to help safeguard sensitive business data in the event of a lost or stolen PC.

PIONEERING THE PRIVATE AI SPACE

Iterate.ai is already looking ahead to optimizing its application for the next generation of AMD processors. Singh highlighted the AMD roadmap and partnership. “In the next 12 to 18 months, the aim is to leverage even bigger LLMs with bigger context windows to satisfy more and more customer use cases,” he said. “I think AMD is doing great in their vision for the future, and we’re excited to be their partner so we can jump on new opportunities and build those use cases.”

AMD Ryzen AI PRO processors help Iterate.ai bring enterprise-scale private AI to the workforce.

WANT TO LEARN HOW AMD PRO PROCESSORS MIGHT WORK FOR YOU?
Sign up to receive our Business content:
www.amd.com/en/preferences/sign-up.html

ABOUT ITERATE.AI

Iterate.ai is an innovation-focused software company that builds enterprise-ready AI platforms and applications for large organizations. Headquartered in San Jose, California, Iterate.ai offers Interplay, a patented low-code AI platform, and private AI solutions such as Generate that help enterprises rapidly build, deploy, and scale secure generative AI experiences across industries including retail, financial services, and technology. For more information visit iterate.ai.

ABOUT AMD

For more than 50 years AMD has driven innovation in high-performance computing, graphics, and visualization technologies. Billions of people, leading Fortune 500 businesses, and cutting-edge scientific research institutions around the world rely on AMD technology daily to improve how they live, work and play. AMD employees are focused on building leadership high-performance and adaptive products that push the boundaries of what is possible. For more information about how AMD is enabling today and inspiring tomorrow, visit the AMD (NASDAQ: AMD) website, blog, LinkedIn, and X pages.

DISCLAIMERS

All performance and cost savings claims are provided by Iterate.ai and have not been independently verified by AMD. Performance and cost benefits are impacted by a variety of variables. Results herein are specific to Iterate.ai and may not be typical. GD-181

The information contained herein is for informational purposes only and is subject to change without notice. While every precaution has been taken in the preparation of this document, it may contain technical inaccuracies, omissions and typographical errors, and AMD is under no obligation to update or otherwise correct this information. Advanced Micro Devices, Inc. makes no representations or warranties with respect to the accuracy or completeness of the contents of this document, and assumes no liability of any kind, including the implied warranties of noninfringement, merchantability or fitness for particular purposes, with respect to the operation or use of AMD hardware, software or other products described herein. No license, including implied or arising by estoppel, to any intellectual property rights is granted by this document. Terms and limitations applicable to the purchase or use of AMD products are as set forth in a signed agreement between the parties or in AMD’s Standard Terms and Conditions of Sale. GD-18u.

COPYRIGHT NOTICE