Startup Pens Generative AI Success Story With NVIDIA NeMo

Using NVIDIA AI software, ‘Writer’ builds LLMs that are helping hundreds of companies create content.

Author: Chintan Patel

Machine learning helped Waseem Alshikh plow through textbooks in college. Now he’s putting generative AI to work, creating content for hundreds of companies.

Born and raised in Syria, Alshikh spoke no English, but he was fluent in software, a talent that served him well when he arrived at college in Lebanon.

“The first day they gave me a stack of textbooks, each one a thousand pages thick, and all of it in English,” he recalled.

So, he wrote a program — a crude but effective statistical classifier that summarized the books — then he studied the summaries.

In 2014, he shared his story with May Habib, an entrepreneur he met while working in Dubai. They agreed to create a startup that could help marketing departments — which are always pressured to do more with less — use machine learning to quickly create copy for their web pages, blogs, ads and more.

“Initially, the tech was not there, until transformer models were announced — that was something we could build on,” said Alshikh, the startup’s CTO.

“We found a few engineers and spent almost six months building our first model, a neural network that barely worked and had about 128 million parameters,” an often-used measure of an AI model’s capability.

Along the way, the young company won some business, changed its name to Writer and connected with NVIDIA.

“Once we got introduced to NVIDIA NeMo , we were able to build industrial-strength models with three, then 20 and now 40 billion parameters, and we’re still scaling,” he said.

NeMo is an application framework that helps companies curate their training datasets, build and customize large language models (LLMs), and run them in production at scale. Organizations everywhere from Korea to Sweden are using it to customize LLMs for their local languages and industries.

“Before NeMo, it took us four and a half months to build a new billion-parameter model. Now we can do it in 16 days — this is mind blowing,” Alshikh said.

In the first six months of this year, the startup’s team of fewer than 20 AI engineers used NeMo to develop 10 models, each with 30 billion parameters or more.

That translates into big opportunities. Hundreds of businesses now use Writer’s models that NeMo customized for finance, healthcare, retail and other vertical markets.

The startup’s customer list includes household names like Deloitte, L’Oreal, Intuit, Uber and many Fortune 500 companies.

Writer’s success with NeMo is just the start of the story. Dozens of other companies have already downloaded NeMo.

The software will be available soon for anyone to use. It’s part of NVIDIA AI Enterprise , full-stack software optimized to accelerate generative AI workloads and backed by enterprise-grade support, security and application programming interface stability.

Some customers run Writer’s models on their own systems or cloud services. Others ask Writer to host the models, or they use Writer’s API.

“Our cloud infrastructure, managed basically by two people, hosts a trillion API calls a month — we’re generating 90,000 words a second,” Alshikh said. “We’re delivering high-quality models that compete with products from companies with larger teams and bigger budgets.”

Writer uses the Triton Inference Server that’s packaged with NeMo to run models in production for its customers. Alshikh reports that Triton, used by many companies running LLMs , enables lower latency and greater throughput than alternative programs.

“This means you can run a service for $20,000, instead of $100,000, so we can invest more in building meaningful features,” he said.

Writer is also a member of NVIDIA Inception , a program that nurtures cutting-edge startups. “Thanks to Inception, we got early access to NeMo and some amazing people who guided us through the process of finding and using the tools we need,” he said.

Now that Writer’s text products are getting traction, Alshikh, who splits his time between homes in Florida and California, is searching the horizon for what’s next. In today’s broad frontier of generative AI, he sees opportunities in images, audio, video, 3D — maybe all of the above.

“We see multimodality as the future,” he said.

Check out this page to get started with NeMo. And learn about the early access program for multimodal NeMo here .

And if you enjoyed this story, let folks on social networks know using the following, a summary suggested by Writer:

“Learn how startup Writer uses NVIDIA NeMo software to generate content for hundreds of companies and rack up impressive revenues with a small staff and budget.”

Original URL: https://blogs.nvidia.com/blog/2023/08/08/writer-nemo-generative-ai/