CASE STUDY: SHOPIFY

Ensuring accurate product attributes at the source

Fine-tuned LLaVA optimized Shopify listings with information extracted from product images

Meta shopify

At a glance

Shopify used a fine-tuned, lightweight LLaVA model to extract information from product images, help merchants create effective listings in real time and optimize performance for the billions of products sold on the platform.

Use case
Extracting metadata from billions of product images and listing text to generate product attributes at scale

Goal
Assist merchants and improve listing search performance

Llama version
Llama 2 7B-based LLaVA model fine-tuned for multimodal reasoning

Deployment
LMdeploy on 100 NVIDIA A100 GPUs

Results*

Tens of billions
of tokens processed every day

Improved
merchant listing experience

More accurate
product attributes and metadata

Conversational search
Customers search with natural language tools

*All results are self-reported and not identifiably repeatable. Generally, expected individual results will differ.

The challenge

Accurately inferring product attributes from images across a self-service platform with millions of stores

Shopify merchants created listings on their own by uploading images, writing copy, assigning product attributes and categorizing the product. With the right categorization and attributes, products show up more accurately in search results. But this manual process left room for error and could result in incomplete attributes or incorrect categorization, which diminished search visibility.

Improving listing accuracy with image-aware AI

Product images are the backbone of e-commerce, and Shopify merchants upload millions of product images to the platform every day. Each product image contains valuable information including the product's color, material, size and more.

As multi-modal models matured, Shopify developers saw a new way to solve the product attribute challenge: use AI agents to extract information from product images and help merchants tag and categorize their products accurately so that they are more discoverable.

shopify

Shopify is an all-in-one commerce platform that powers millions of businesses worldwide. By making it easier to start and grow a business, Shopify helps people around the world achieve independence through commerce.

- **Industry:** Commerce
- **Company size:** Millions of merchants across 175+ countries, generating nearly US$9 billion in annual revenue

The solution

Dual agents for real-time and offline attribute extraction and product categorization

Faced with millions of new product images per day, billions of existing images and more than 10,000 possible attribute categories, Shopify developers realized they needed to split the solution into two use cases.

The first use case was a real-time AI agent that helped merchants create accurate, optimized product attributes as they entered new product listings. The second use case was a high-volume, offline agent that could extract attributes from the billions of images and product descriptions on the Shopify platform and create updated listings that would perform better in search.

> "Closed models performed well, but their pricing structures made them prohibitively expensive. When you need to run billions of multi-modal inferences a day, per-token charges aren't sustainable."
>
> **Kshetrajna Raghavan,** Machine Learning Leader, Shopify

Open-source LLaVA delivered the right mix of capabilities and low cost

Both the real-time and offline solutions needed to extract semantic context from product images and text, identify basics like product size and merchandising categories and then generate more nuanced, complex metadata for image descriptions and power ratings. All AI-generated metadata had to be accurate and consistent to minimize errors and maximize performance in search and discovery. Finally, the AI agents needed to detect and flag images that violated Shopify Terms of Service.

Developers began prototyping with closed, proprietary models, but the high per-inference costs made them impractical for processing billions of tokens per day. Llama-based Large Language and Vision Assistant (LLaVA) — an open-source multi-modal model that combines a vision encoder with Llama 2 — delivered competitive results without per-token charges, making it an ideal choice for production inference at a massive scale.

Creating training data took AI and human oversight

The Shopify team needed higher performance than stock LlaVA could deliver, so developers decided to retrain the model. They began by building a training dataset from existing merchant-created image entries.

The resulting merchant dataset was relatively modest and contained occasional mistakes. To create a more robust dataset, the team used a leading closed-source model as a teaching model to further categorize images and generate additional metadata. Finally, human evaluators vetted the output to ensure the training dataset was richly tagged and accurately cataloged.

Quantized training saved time and reduced computing costs

Training began with a Llama 2 13B-based LLaVA model, but a Llama 2 7B-based model showed equivalent performance after initial training, so the team chose the smaller model for development.

The team leveraged Quantized Low-Rank Adaptation (QLoRA) for the first training rounds. QLoRA used quantization and selective parameter updating to reduce memory footprints and compute workloads so training could run on more modest hardware.

Training Llama 7B at lower precision with QLoRA achieved predictions within 95% of the required accuracy. To reach their final target, the team switched to full-precision training on bare metal resources in the Google Cloud Platform.

> "We were pleasantly surprised to find that with fine-tuning, the smaller Llama 7B model delivered the accuracy we needed. Using a smaller parameter size sped up training and saves compute cycles in production."
>
> **Agatha Krajewski,**
> Director of Product, Shopify

> "Analyzing every image on the platform and generating metadata is an immense task. Using smaller-parameter Llama models makes it possible without expensive, bleeding-edge hardware."
>
> **Kshetrajna Raghavan,**
> Machine Learning Leader, Shopify

Optimizing production for massive-volume inference

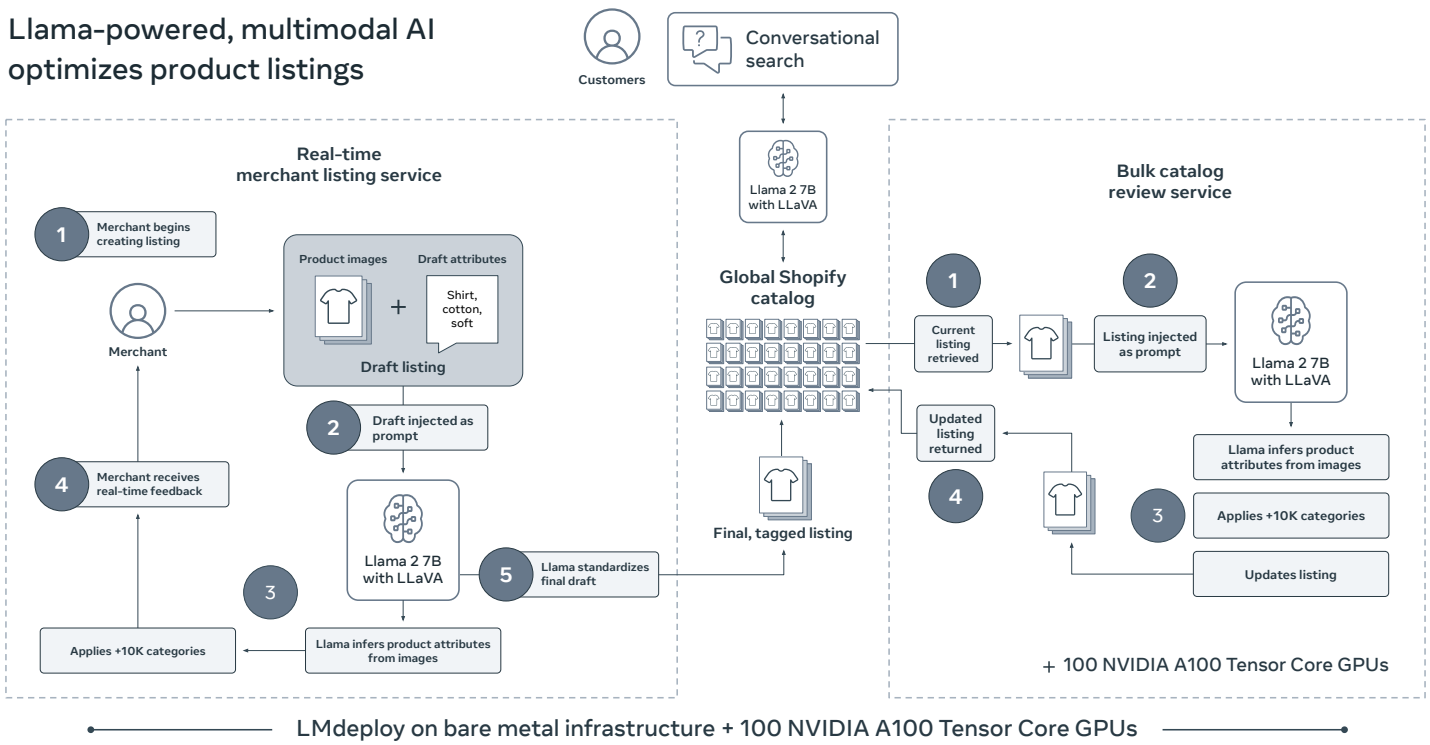
Given the scale of inference tasks the solution needed to process, the Shopify team did everything possible to reduce computing costs. They chose LMdeploy — a toolkit for compressing, deploying and serving large language models — to increase production efficiency.

LMdeploy managed compute resources and applied Activation-aware Weight Quantization (AWQ) automatically. This ran the most important weights in the model at full precision and quantized less important functions to reduce memory and compute demands.

Llama 2 7B scaled up while keeping inference costs low

Carefully managing resources and workload footprints at runtime helped the Llama 2 7B-powered solution deliver high-performance AI while minimizing computing costs. Still, extracting data and tagging the constantly growing Shopify image library — a workload well over tens of billions of tokens per day — required 100 NVIDIA A100 Tensor Core GPUs.

Once the complete solution was running efficiently, the team deployed it to the Shopify platform where it began coaching merchants and working through existing images.



Shopify used Llama 2 7B to extract product attributes from images, create uniform metadata, categorize listings and provide conversational search services for customers.

The outcome

Happier merchants, more accurate product attributes

The Llama-powered merchant assistant provided a conversational experience that made it far easier to create a product listing and ensured every new image had accurate metadata. The offline solution delivered platform-wide improvements as it reviewed billions of images, improved descriptions and added nuance by applying more than 10,000 possible metadata categories.

The improved tags benefited shopper discovery, SEO and site performance. Shoppers found a broader range of relevant products with simple natural language queries. For merchants, their products were better surfaced to the most interested shoppers, increasing their sales. Since the system was largely automated and entirely open source, it transformed Shopify's metadata without expensive human evaluators or per-token inference fees.

Tens of billions
of tokens processed every day

Improved
merchant listing experience

More accurate
product attributes and metadata

Conversational search
Customers search with natural language tools

"Without Llama-powered, open-source LLaVA, providing real-time listing support for merchants would be prohibitively expensive — let alone optimizing Shopify's entire product catalog. Llama 2 7B's small footprint and zero per-token fees made our project both technically and economically feasible."

Agatha Krajewski,
Director of Product, Shopify

Conclusion

Specialized Llama vision models are on the horizon

The Shopify team has already tested Llama 3.2 11B Vision. They find it works well off the shelf and expect it will perform even better after fine-tuning.

How can Llama help your business?

See how open-source Llama brings unmatched control, customization and flexibility to generative AI application development and deployment.

Learn More

Related Stories ▸

Meta