CASE STUDY: AITEM SOLUTIONS

Equipping veterinarians to make faster, better clinical decisions

Llama-enabled AI assistant designed for veterinary care delivers specialized expertise in complex, high-stakes care moments

Meta | AITEM

At a glance

AITEM Solutions, a member of the NVIDIA Inception program, developed LAIKA — the first AI assistant for veterinary care — using Llama 3.1 70B Instruct. This generative AI solution uses custom retrieval-augmented generation (RAG) pipelines to deliver accurate, domain-specific responses to complex care questions. Designed for busy, high-pressure environments, LAIKA helps veterinarians make faster, more confident decisions while reducing mental load.

Industry
Technology

Use case
Improving veterinary care with a specialized AI assistant

Goal
Make it easier and faster for vets to navigate care successfully

Llama version
Llama 3.1 70B Instruct

Deployment
Hosted cloud

Results*

75x
faster than human support at analyzing critical cases

5-10
minutes saved for every analysis

17.5x
cheaper than human support

4 stars
out of 5 — rated by clinical vets

*All results are self-reported and reflect the partner's individual circumstances and performance. Others' results and experiences may vary.

The challenge

Overwhelmed veterinarians with limited support

Like human healthcare, veterinary care is a complex and emotionally demanding field. Veterinarians make critical decisions under pressure — with little margin for error — that affect the lives of animals and pets that people care deeply about.

But unlike human healthcare, vets are expected to understand a wide variety of species and breeds with unique physiological behaviors and traits. That makes specialized care difficult. They often work independently, without the support of nurses, specialists or assistants — leaving the full weight of responsibility on the individual vet. Because veterinary services have little to no insurance support, there is limited funding and investment in the sector, while demand for care grows. These pressures make it increasingly difficult for vets to consistently deliver efficient, high-quality care.

Improving clinical decision-making and lowering stress

The team at AITEM Solutions was on a mission to create a domain-specific AI assistant for the complex world of veterinary medicine. The solution needed to help vets tackle daily activities while under intense time pressure, handle the variability of cases and overcome limited access to specialists. Ultimately, the team wanted to enable faster and more effective clinical decision-making while reducing stress and mental load for vets.

AITEM Solutions develops cutting-edge AI products with a strong focus on digital healthcare. The company's solutions enable veterinary medicine, medical imaging, diagnostic support, proteomics, data analysis and more — all to enhance the precision and efficiency of modern medicine and drive human decisions toward zero errors.

- **Industry:** Technology
- **Company size:** 18 employees

> "The answer wasn't just another AI chatbot or general-purpose assistant. The industry needed a veterinary-specific AI assistant uniquely built to support them in their everyday work."

Davide Tricarico,
CTO, AITEM Solutions

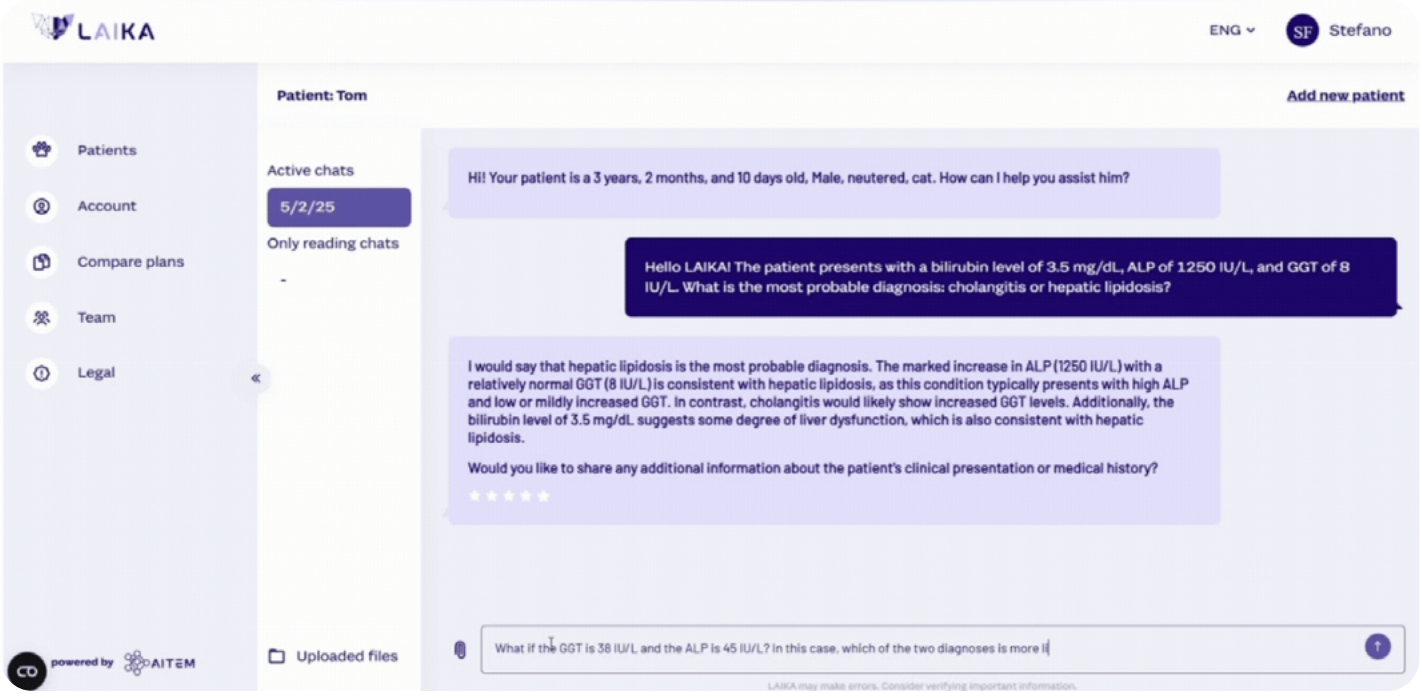
The solution

A domain-specific veterinary AI assistant backed by Llama

AITEM Solutions created LAIKA, the first AI assistant specifically designed for veterinary care. They first used Llama 2 and continued to upgrade as models were released to Llama 3.1 70B Instruct. LAIKA acts as a specialist colleague, offering expertise for clinical cases beyond a vet's immediate knowledge or a second opinion when a vet wants to double-check their assessment. Additionally, LAIKA can help with communication by guiding vets on how to talk to pet owners during emotionally challenging moments.

Incorporating Llama models into the AI assistant enabled LAIKA to deliver fast and accurate responses during vets' daily clinical work. By retrieving from a specialized RAG pipeline that incorporates the NVIDIA NeMo Retriever reranking model, Llama allows LAIKA to take into account relevant medical literature and patient-specific factors. This helps ensure that vets consider all aspects of a patient's situation and reduce missteps in care.

> "LAIKA's overall goal is to provide fast, accurate and specialized support for veterinarians in their clinical decision-making. Llama is a cornerstone to achieving that goal."
>
> Davide Tricarico,
> CTO, AITEM Solutions



LAIKA offers easy access to specialized expertise when vets need it most.

Llama stands out with its openness and flexibility

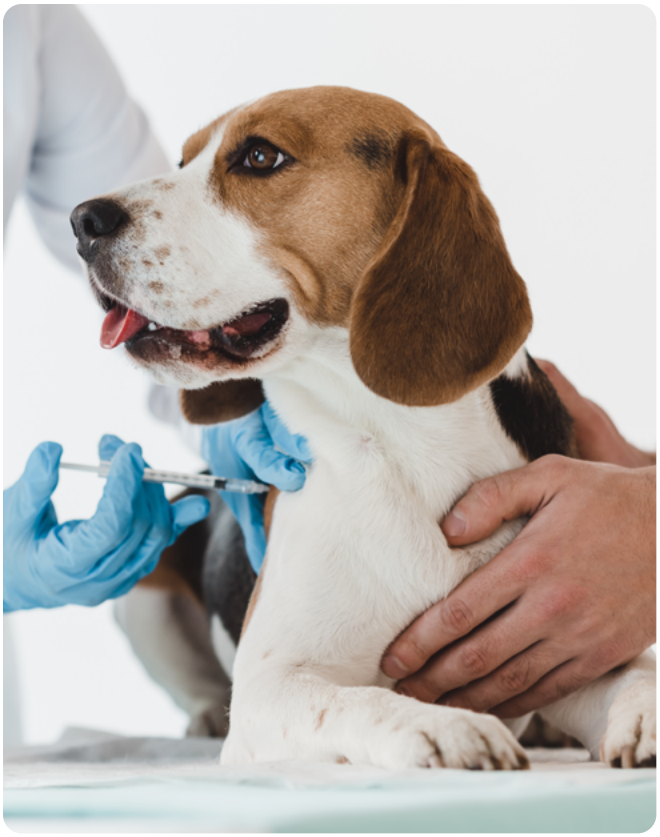
After an iterative evaluation process involving AITEM Solutions' internal dataset and real-world feedback from veterinarians, the team chose Llama as a foundation model for LAIKA. Not only did Llama provide the best performance, including quality of responses and speed, but it also delivered on the critical benefits of openness.

The open weights gave AITEM Solutions the flexibility to fine-tune and customize the model for veterinary use cases. Beyond that, they were able to select the right version for LAIKA thanks to the availability of Llama in different formats. The team could work with the raw model, a containerized NVIDIA Inference Microservice (NIM) or a serverless API from different providers, helping to avoid vendor lock-in.

NVIDIA NIM can streamline the design of NLP pipelines using large language models (LLMs). These microservices simplify the deployment of generative AI models across systems, allowing teams to self-host LLMs while offering standard APIs to build applications. NIM abstracts model inference internals like execution engines and runtime operations, ensuring optimal performance with TensorRT-LLM, vLLM and others. AITEM developers selected this modality to deploy the NeMo Retriever reranking model as part of the RAG pipeline in production.

To minimize sensitive data exposure, AITEM hosted open-source Llama on company-controlled infrastructure. Docker containers running on AWS and Scaleway provided flexible, secure deployments similar to on-premises hosting. Llama's commercial use license and active developer community were additional advantages, giving AITEM access to collective expertise and best practices that accelerated development, reduced errors and informed better technical decisions.

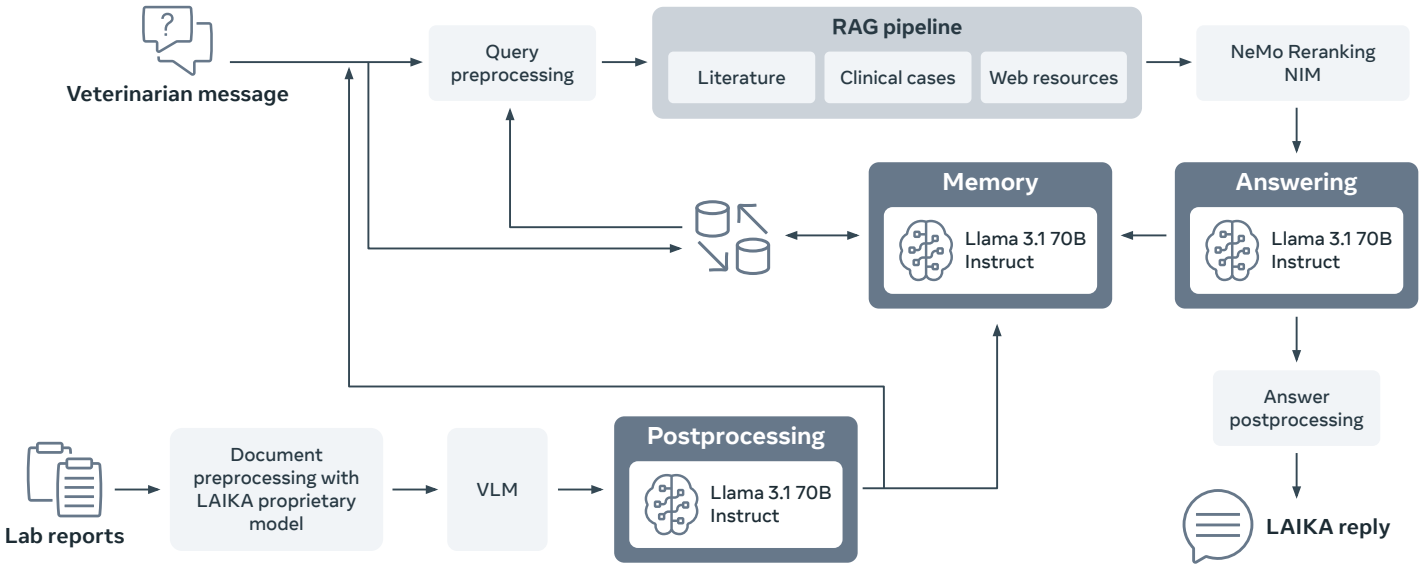
> "Having access to a large and active open-source community had a real impact. It allowed us to identify and overcome issues faster using best practices shared by others who were working on similar problems."
>
> Davide Tricarico,
> CTO, AITEM Solutions



Providing accurate, context-aware answers with Llama and RAG

In developing the LAIKA AI assistant, AITEM Solutions created a broad architecture that included RAG pipelines, deterministic models, fine-tuned vision components, vision-language models and other LLMs. The Llama model was brought on to help solve an important problem for LAIKA: delivering response quality and accuracy for a highly complex and specialized field.

In reality, out-of-the-box foundation models had limited exposure to veterinary science data and lacked clinical reasoning skills. With a fine-tuned Llama model for generative AI and a custom RAG pipeline, LAIKA could retrieve information from veterinary-specific knowledge bases and reproduce the typical reasoning process vets use for diagnostics. That meant LAIKA gave more accurate, context-aware answers that aligned with the needs of real vets.



This implementation leverages multiple instances of Llama 3.1 70B Instruct and incorporates a reranking model developed by NVIDIA, which is available as a NIM container in NeMo Retriever.

Prompt engineering and testing optimize responses

Busy vets working in stressful environments don't have time to carefully phrase their prompts and questions. LAIKA needed to be able to interpret and respond well to messy, unstructured input with specific, trustworthy insights. At the same time, answers must be in the right tone — as if talking to a colleague — with the right level of detail.

During LAIKA's beta phase, AITEM Solutions used prompt engineering with one- to few-shot learning to regulate LAIKA's tone of voice and deliver useful text structures. Then, the team worked closely with over 100 vets to collect continuous feedback and refine the way LAIKA communicates, adopting specialist vocabulary and right-sizing the amount of information in responses.

They allowed vets to rate LAIKA responses on a scale of one to five stars. In the beginning of the beta phase with Llama 2, only 23% of responses received four stars or higher. After refining the RAG pipeline, that score was raised to 33%. With additional advances to LAIKA's architecture and the shift to Llama 3, the AI assistant saw major improvements. 68% of new ratings were four stars or higher. The last RAG pipeline version and upgrade to Llama 3.1 70B Instruct saw the best results yet: 70% of responses were rated four stars or more, with over 50% getting full five-star ratings.

The outcome

Empowering vets to deliver better care faster — and live better lives

The LAIKA AI assistant, backed by Llama, helps vets navigate the complex and high-pressure world of veterinary care with more success and less stress. Vets can count on LAIKA to help them speed up clinical decision-making with easy access to specialized expertise. In analyzing cases, LAIKA is 75x faster than human support, which saves vets 5-10 minutes per analysis. From an operational cost perspective, LAIKA is 17.5x cheaper than human support.

This new level of productivity is in part due to how LAIKA bolsters vets' confidence. They no longer feel like they are acting alone in care situations with heightened emotional stakes. And they spend less time wavering between multiple options when facing uncertainty. Using LAIKA reduces their mental load, freeing them to deliver better care to more animals and pets and improve outcomes.

Developed in close collaboration with veterinarians, LAIKA is built for real-world practice. Its responses consistently earn an average rating of four out of five stars for performance across daily clinical cases, reflecting its precision, practicality and value in the field.

> "The Llama model family has played an important role in enabling these results. Not just because of its performance, but because of its open nature, flexibility and community support — all of which make it possible to build domain-specific, real-world AI tools."
>
> Davide Tricarico,
> CTO, AITEM Solutions

75x
faster than human support at analyzing critical cases

5-10
minutes saved for every analysis

17.5x
cheaper than human support

4 stars
out of 5 — rated by clinical vets

Conclusion

Unlocking specialized AI support in medicine and beyond

After developing LAIKA, AITEM Solutions saw firsthand that AI delivers the greatest impact when grounded in real-world use, supported by the right tools and developed in close collaboration with the professionals it serves. The team plans to take these learnings and apply them to more use cases. As for LAIKA, AITEM Solutions plans to continue to advance performance for the AI assistant with each new Llama release.

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 ▸

© 2025 META

Meta