CASE STUDY: ENTERPRISE CONSULTING PARTNERS

Building better AI agents for nuanced knowledge work

Llama and LoRA bring expert help to automate complex tasks

Meta

At a glance

Enterprise Consulting Partners (ECP) worked with Predibase to enhance its AI agent with Llama 3.1 8B Instruct and low-rank adaptation (LoRA). The upgrade improves semantic understanding and automates enterprise-scale knowledge tasks while using only a fraction of the computing power required by proprietary AI solutions.

Use case
Improve existing AI assistant's semantic understanding, accuracy and speed

Goal
Transform information gathering and synthesis

Llama version
Llama 3.1 8B Instruct with LoRA

Deployment
Predibase

Results*

25M
annual queries

7%
more accurate than GPT-4o mini after fine-tuning

4-second
round trip time

1M
hours saved

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

The challenge

A confused AI agent was misunderstanding prompts and delivering incorrect responses

In 2023, shortly after ChatGPT debuted, ECP's 5,000-member technology team launched a generative AI assistant for the entire organization. Powered by GPT-3.5, the AI assistant transformed how the firm's analysts and knowledge workers navigated its extensive knowledge bases.

As the tool matured from chat client to researcher and writing assistant, it began to struggle with nuanced tasks. For example, when users asked for help drafting an email, it returned documents with the keyword "email" rather than launching an email application and generating content.

Improving accuracy and semantic understanding

Prompt engineering and retrieval augmented generation (RAG) worked well for general-knowledge chat services, but accuracy declined once the AI assistant was integrated into enterprise platforms like Microsoft SharePoint and media asset systems. Although the team explored fine-tuning expert models for each workflow, training costs and increased system complexity proved to be significant obstacles.

Enterprise Consulting Partners (ECP) is a pseudonym for a leading professional services firm specializing in risk, strategy and people. The firm helps corporate and public sector leaders navigate an increasingly dynamic environment by addressing the most complex challenges of our time.

- **Industry:** Professional services
- **Company size:** 90,000+ global employees, $24 billion annual revenue

"It wasn't just about the raw expense. Fine-tuning introduces operational complexities. Splitting models by audience or department multiplies infrastructure and complexity. Once specific data is in a model, how do we ensure the right people have the right access?"

Global Chief Information Officer
Enterprise Consulting Partners

The solution

Retool the AI assistant with Llama, LoRA and LoRAX

The ECP team partnered with Predibase to explore LoRA as an alternative to full-weight model-tuning and LoRA Exchange (LoRAX) for serving fine-tuned adaptations at enterprise scale. The team chose Llama 3.1 8B Instruct as the foundation model for its small size and low-latency performance.

ECP's AI agent now brings conversational AI assistance to unstructured institutional knowledge, current research and enterprise tools. Llama's superior semantic understanding, multimodal skills and 128K token window are major upgrades to the AI assistant's base capabilities, while LoRA adapters deliver task- and tool-specific expertise.

Fine-tuning helped Llama 3.1 8B Instruct best GPT-3.5 and GPT-4o mini

After fine-tuning, Llama 3.1 8B Instruct was 10-12% more accurate than GPT-3.5, despite having roughly 25x fewer parameters. In tests against similarly sized GPT-4o mini, Llama 3.1 8B Instruct proved 7% more accurate. In production, the agentic system powered by fine-tuned Llama consistently met or exceeded GPT-3.5's sub-four-second round-trip response times.

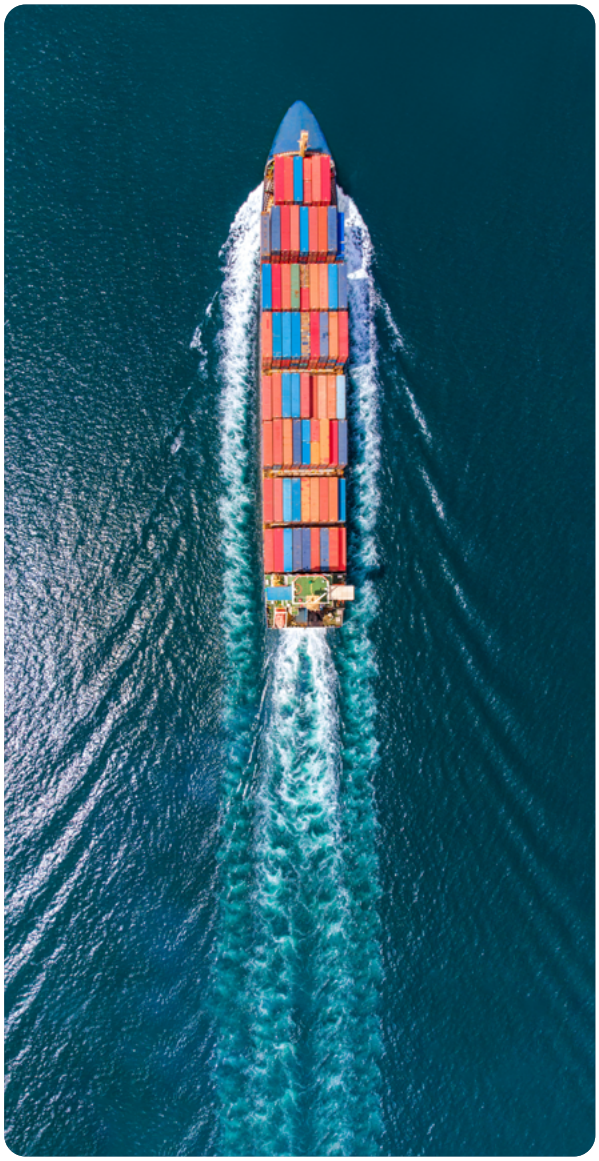
> "Llama's impressive speed and ability to efficiently manage enterprise-scale queries has made our AI assistant more responsive while LoRA fine-tuning is increasing its understanding, expertise and accuracy. Our upgraded AI assistant will deliver rapid access to vast stores of institutional knowledge and take on mundane tasks, improving overall operational efficiency."
>
> Global Chief Information Officer
> Enterprise Consulting Partners

ECP uses LoRA to develop fine-tuned expertise with less training effort

With fine-tuning, smaller models like Llama 3.1 8B Instruct can exceed the performance of large, general-purpose models like GPT — which can have 10x or 100x the number of parameters. However, fully retraining smaller models is still a time-consuming and expensive process.

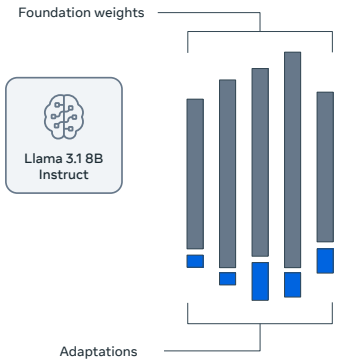
ECP created fine-tuned performance using the Predibase environment and LoRA, which uses small adaptations to a model's weights to reduce fine-tuning time 5x to 10x. LoRA made it possible for the team to iterate quickly and experiment with multiple task-specific adapters to see how they improve response quality.

The Predibase production environment will also streamline model serving with LoRAX, a solution that loads adapters at runtime. Instead of launching multiple instances of the entire model to serve each request, LoRAX runs a single Llama 3.1 8B Instruct instance and launches LoRA adaptations on demand. LoRAX can spin up multiple fine-tuned LoRA adaptations simultaneously in the same basic computing footprint of Llama 3.1 8B Instruct.

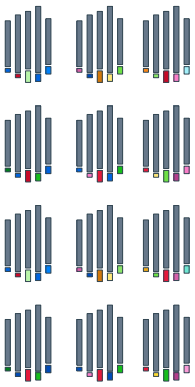


Deploying dozens of fine-tuned Llama models for the price of one

1. Use LoRA to create a fine-tuned adaptation of Llama weights for specific tasks.

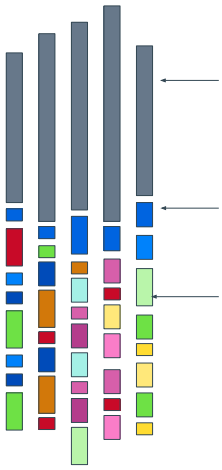


2. Create adaptations for as many specific tasks as the use cases require.



Create fine-tuned adaptations with LoRA on Predibase

3. Serve a single model and load adaptations on demand at runtime using LoRA Exchange (LoRAX).



Serve adaptations with LoRAX on Predibase

ECP is developing fine-tuned performance using LoRA. In the future, the production pipeline will be able to deploy dozens of fine-tuned adaptations using a single instance of Llama.

"With Llama and Predibase, the entry price and complexity dropped. Suddenly, we didn't need separate infrastructure for every fine-tuned model. Training became incredibly cost-effective — tens of dollars, not hundreds of thousands. This unlocked a new wave of automation use cases that previously weren't economical."

Global Chief Information Officer
Enterprise Consulting Partners

The outcome

Expert AI agent delivers nuanced help, saves teams over one million hours

The retooled AI assistant is a solid success. The combination of Llama and fine-tuned LoRA adaptations increased the assistant's ability to comprehend nuanced requests, improved answer accuracy and expanded its capabilities.

Most importantly, the AI assistant has shouldered an astonishing amount of knowledge work, freeing ECP teams to focus on higher-order thinking and spend more time serving clients. By answering more than 25 million cumulative queries across the organization in its first year, the AI assistant has saved at least one million hours of team time and helped create a massive increase in productivity.

25M
annual queries

7%
more accurate than GPT-4o mini after fine-tuning

4-second
round trip time

1M
hours saved

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

"Our journey with generative AI started before it became a mainstream tool. With Llama, we upgraded to a fast, accurate, expert tool that upwards of 85,000 global employees tap into every day."

Global Chief Information Officer
Enterprise Consulting Partners

Conclusion

Sharing the power of AI enterprise-wide

While the AI assistant is busy helping ECP put decades of accrued knowledge and the organization's latest insights to work, the technology team is busy expanding capabilities throughout the company.

The roadmap includes advanced document classification systems, enhanced extraction capabilities and new specialized use cases. Development teams across ECP have free access to the AI assistant's technology, including Llama and LoRA fine-tuning, so they can tailor it to their own needs and invent new use cases that centralized teams might never imagine.

> "We've put Llama-powered AI into our employees' hands. Any development team across the organization can experiment with our AI assistant technology, apply it to new use cases and invent new tools. The success we've seen so far will only grow over time."
>
> Global Chief Information Officer
> Enterprise Consulting Partners

How can Llama enable 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