Predibase

Beyond the Buzz: A Look at Large Language Models in Production

Survey report on the top trends, challenges, and recommendations for enterprises building with LLMs

Introduction

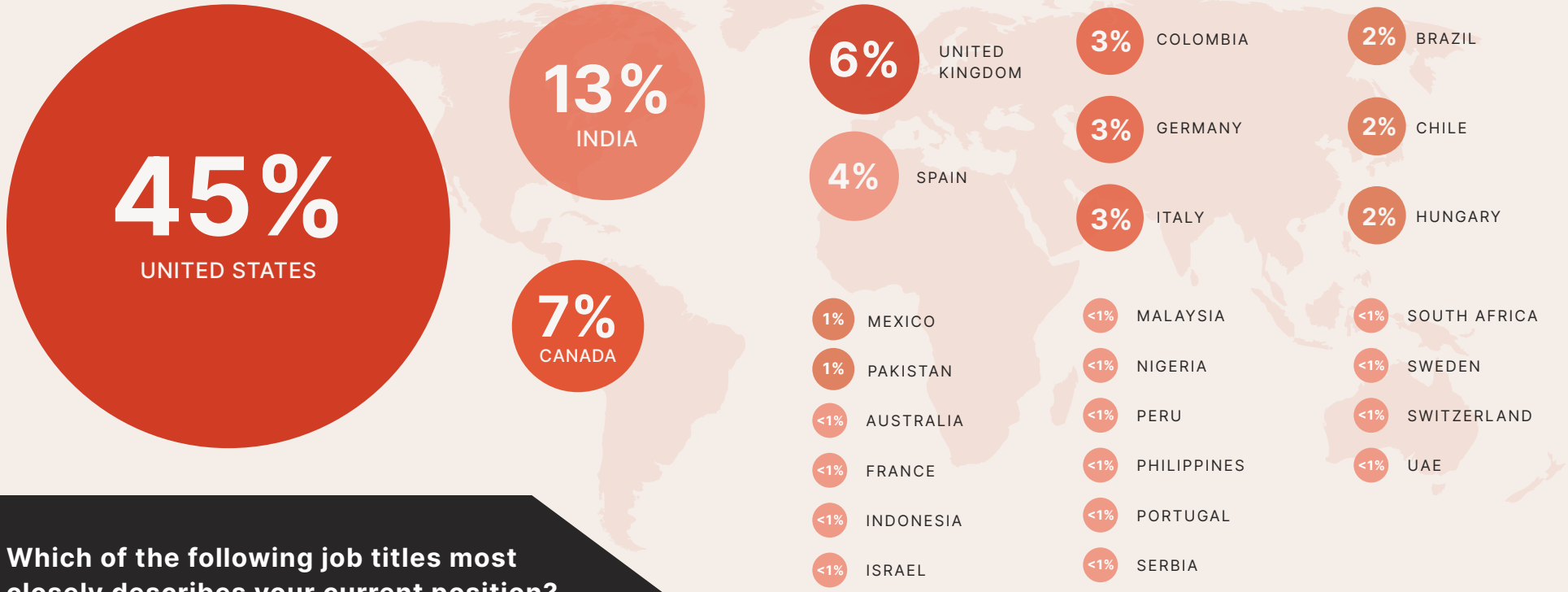
Propelled by the launch of OpenAI's ChatGPT last year, the Large Language Model (LLM) hype train has reached full steam. A new LLM company is born daily, and a bigger and better open-source model is released weekly. With all these options, enterprises are scrambling to determine how to gain a competitive edge with these new AI capabilities. But with all this excitement, the question remains: is this all just hype, and more importantly, how can organizations successfully put LLMs into production?

To better understand how organizations are adopting LLMs for production applications, we surveyed **150** executives, data scientists, machine learning engineers, developers, and product managers at both large and small enterprises. The survey was conducted from May to July of 2023 with respondents representing **29** countries.

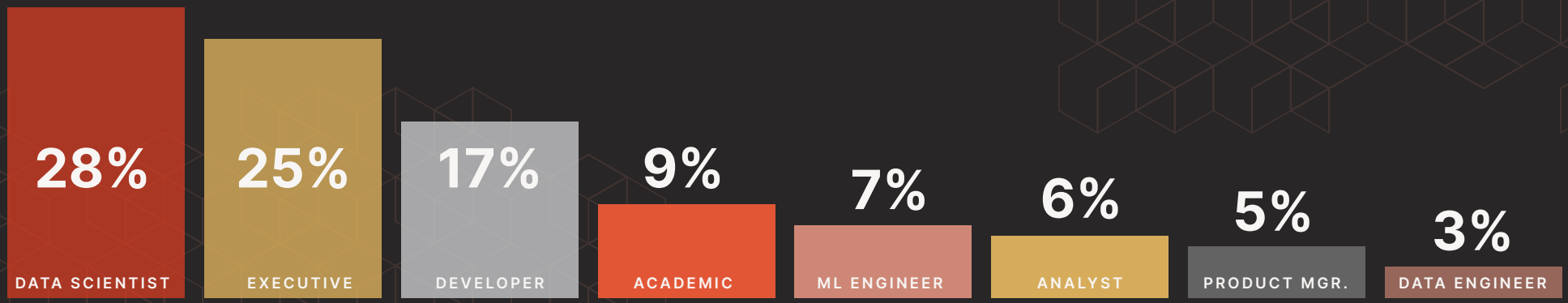
In the survey, we explored motivations for investing in LLMs, key challenges faced when deploying into production, the growing opportunity for open-source LLMs, methods for customization, and the exciting range of LLM-powered use cases. By capturing the diverse opinions of industry professionals, we aim to provide insights into the current landscape of LLM adoption and the opportunities these models present.

Survey Demographics

What country are you located in?



Which of the following job titles most closely describes your current position?



The growth of LLMs and evolving ML workflows

The recent growth of LLMs has been remarkable, revolutionizing ML workflows and opening up new perceptions of what is achievable with AI. With their ability to generate coherent and contextually relevant text, LLMs have rapidly gained widespread adoption, propelling innovation across industries. In fact, AI adoption has more than doubled over the past five years, and as LLMs — such as GPT-4, LLaMa 2, and Flan T5 — have showcased their ability to understand and generate human-like text, the interest in AI-powered applications is skyrocketing (ChatGPT reached 100 million monthly active users just two months after launch).

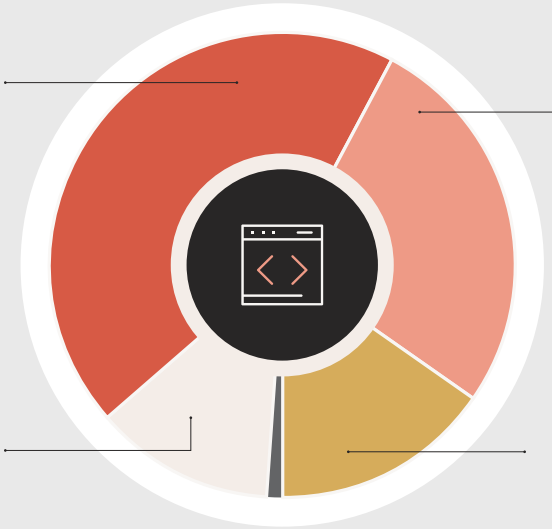
Additionally, their ease of use empowers less-experienced ML practitioners and engineers to rapidly build and prototype AI-powered applications without the labor-intensive process of training models from scratch. This shift in ML workflows allows broader AI adoption across personas within the organization. More experienced practitioners can now focus on finetuning and customizing their models for specific tasks or domains vs. starting from square one.

With **over half (58%)** of data scientists and engineers sharing that they are actively using large language models, it begs the question, what does the future hold for this newfound collaboration between humans and machines?

Enterprise adoption of LLMs

Overall, LLM adoption in the enterprise has shown promise as teams quickly recognized the opportunity and broad set of use cases for LLMs since the advent of ChatGPT in November of last year. Less than 40% of the companies we interviewed have yet to begin working with LLMs; of that group, only 15% have no immediate plans. The large majority are either experimenting or putting LLMs into production. With LLMs still in their infancy, it's no surprise that the largest group of respondents — roughly 50% — are currently in the experimentation phase.

Are you using LLMs at your organization today?



GENERATIVE JOBS ON THE RISE AS IS THE NEED FOR THE RIGHT TECH STACK

OMDIA

According to independent research authority, Omdia (an Informa Tech company), internal experimentation is rapidly converting into actionable investment with jobs specific to generative AI growing at a propitious rate.

Looking at global job requisitions for AI professionals between January 2022 and July 2023, generative AI job openings grew to **2.4%** of all measured AI vacancies. In just one quarter, between April and July, that number grew by a whopping **712%**.[1]

> **Clearly, companies are investing in the personnel and technologies necessary to work with emerging generative AI technologies in support of production-scale outcomes. The trick, of course, will rest not in building these outcomes, but ensuring that they deliver consistent, secure, responsible outcomes. With an increasing desire to customize and deploy open-source models, enterprises will need to invest in operational tooling and infrastructure capable of keeping up with the rapid pace of innovation in the open-source community.**
>
> **BRADLEY SHIMMIN**
> Chief Analyst AI platforms, analytics and data management at Omdia

[1] *https://omdia.tech.informa.com/OM032398/AI-Skills-Tracker--1H23-Analysis*

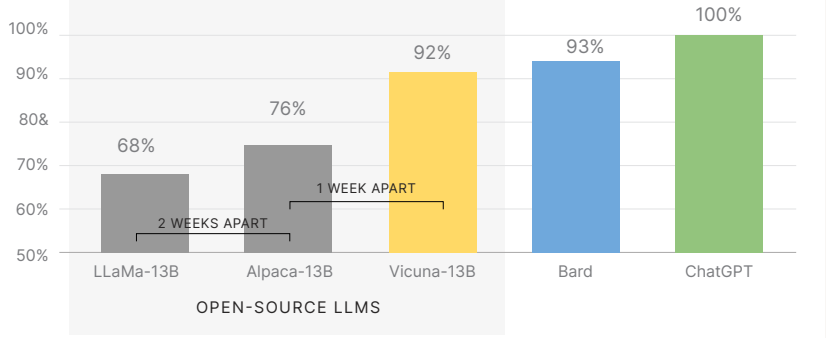
Concerns with Commercial LLMs

Despite the enthusiasm, enterprises are slow to adopt commercial LLMs — like GPT provided by OpenAI — as they share several concerns. In fact, less than a 1/4th of surveyed companies are comfortable using commercial LLMs in production. At a high level, data privacy concerns top the list. In our discussions, nearly 40% of companies voiced concerns about sharing proprietary or sensitive data with LLM vendors. So if "renting" a commercial LLM isn't the answer, what should an enterprise do?

Fortunately, the open-source community is rapidly releasing new and increasingly powerful LLMs that rival commercial offerings almost weekly. Even commercial LLM providers are jumping on the bandwagon. For example, Meta moved away from building closed-source LLMs like LLaMA-1 with their most recent release, LLaMA-2, which is open-source. With so much innovation in the community, we expect companies to continue to adopt open-source models.

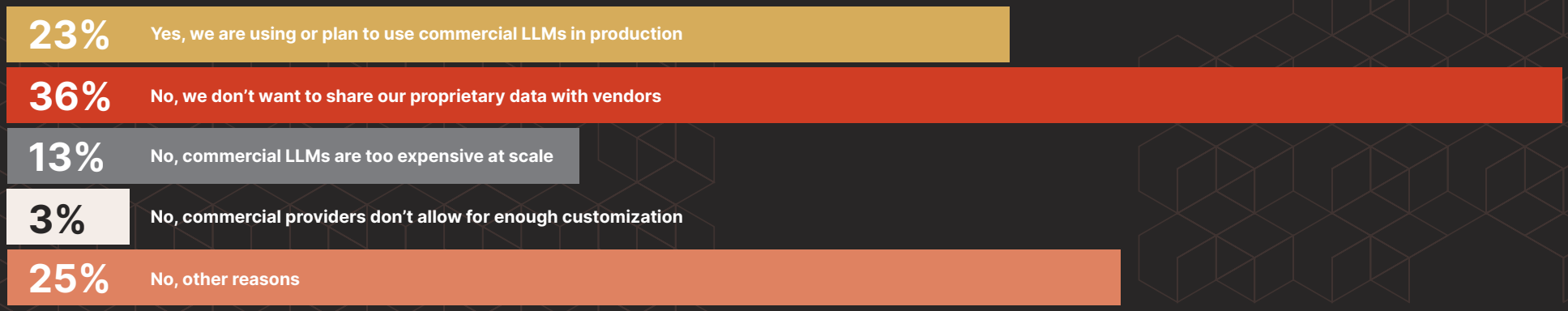
OPEN-SOURCE LARGE LANGUAGE MODELS ARE CATCHING UP

A recently leaked memo from Google highlighted how open-source LLMs are rapidly closing the gap with commercial LLMs.



*GPT-4 grades LLM outputs. Source: vicuna.lmsys.org/

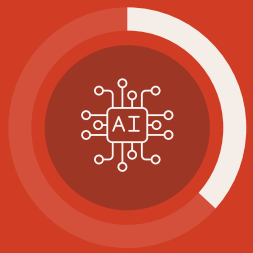
Are you using or planning to use commercial LLMs in production?



Motivations for Investing in LLMs

LLMs offer a broad set of benefits driving these record-level investments. The desire to build generative AI capabilities tops the list but is not the only value driver. Businesses are also interested in leveraging LLMs to accelerate AI projects and tackle new AI/ML use cases that previously were out of reach, especially if they lack deep data science expertise.

What is your primary motivation for investing in LLMs?



37%

DESIRE TO BUILD GENERATIVE AI CAPABILITIES

26%

ACCELERATE THE DEVELOPMENT OF AI/ML PROJECTS

26%

MAKE IT EASIER TO TACKLE NEW AI/ML PROJECTS THAT WERE PREVIOUSLY TOO COMPLEX

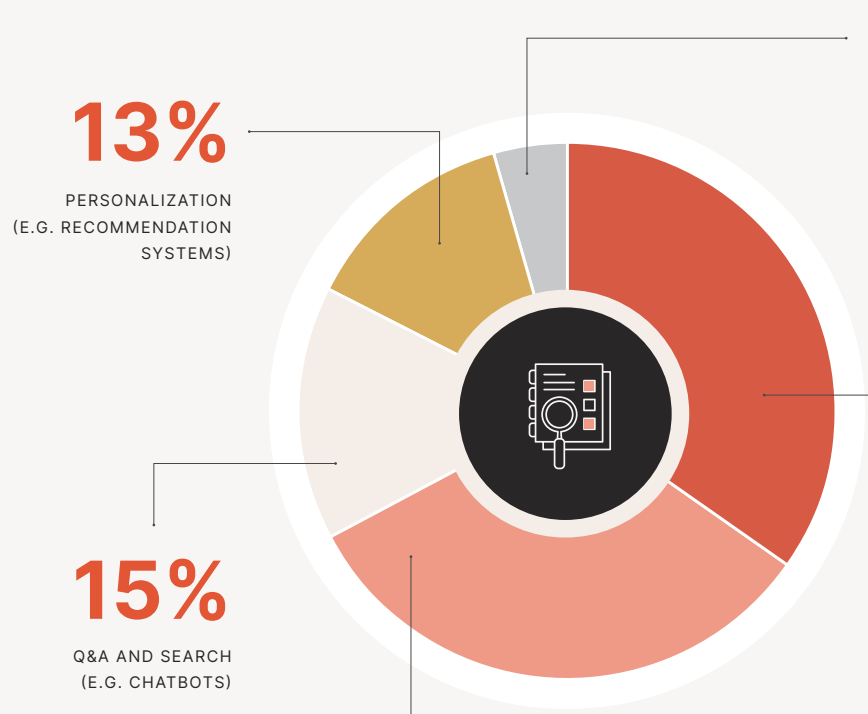
11%

ENABLE A BROADER SET OF PERSONAS AT MY ORGANIZATION TO WORK WITH AI/ML

Exploring the top use cases for LLMs

When it comes to using LLMs in an organization, most minds immediately go to generative use cases and chatbots, thanks to the popularity of ChatGPT. So it's no surprise to see that represented by our data. However, LLMs have a much broader range of applications. An emerging and popular use case is Information Extraction (the next largest response in our survey), which involves leveraging LLMs to convert unstructured data like PDF documents, customer emails, or website content into structured tables for aggregate analytics.

What is your top use case for LLMs?



Exploring the top use cases for LLMs

THE FUTURE OF LLM APPLICATIONS

The use cases below exemplify the remarkable potential for LLMs beyond generative tasks, empowering organizations to extract valuable insights from unstructured data and make informed decisions.

GENERATIVE AI

LLMs can be used for many generative taks. They can be applied to marketing or sales content development, generating persuasive copy for campaigns, or crafting product descriptions. They can also assist in writing code for software development by providing code suggestions or auto-completion. LLMs can also summarize and categorize customer support issues, helping streamline the ticketing process and improve response times.

INFORMATION EXTRACTION

Frequently, teams use LLMs to ask open-ended questions on data, but a more effective approach when calculating aggregate statistics is first to use an LLM to extract unstructured text from documents like PDFs and transform it into structured tables for further analysis. For instance, imagine being an investment bank that wants to utilize LLMs to convert a large corpus of unstructured financial reports (like 10Ks and investor call transcripts) into a structured database containing key metrics for potential investments, such as risk factors, revenues, and number of new customers, locations or products. You could then use SQL queries to aggregate and analyze those metrics for deeper analysis. Another example of Information Extraction is enriching patient data by extracting structured information from doctors' notes or lab reports.

TEXT CLASSIFICATION

Organizations can use LLMs for a broad range of text classification use cases like customer feedback sentiment analysis, understanding customer satisfaction levels, and identifying product areas for improvement. They are valuable in content moderation, flagging inappropriate or sensitive content in user-generated submissions.

Q&A AND SEARCH

LLMs can power customer chatbots, providing accurate and relevant responses to user queries in real time. They can enhance website search engines, enabling users to find desired information quickly and effortlessly. LLMs can also assist in search, helping researchers locate specific information within lengthy documents or academic papers.

PERSONALIZATION/ RECOMMENDER SYSTEMS

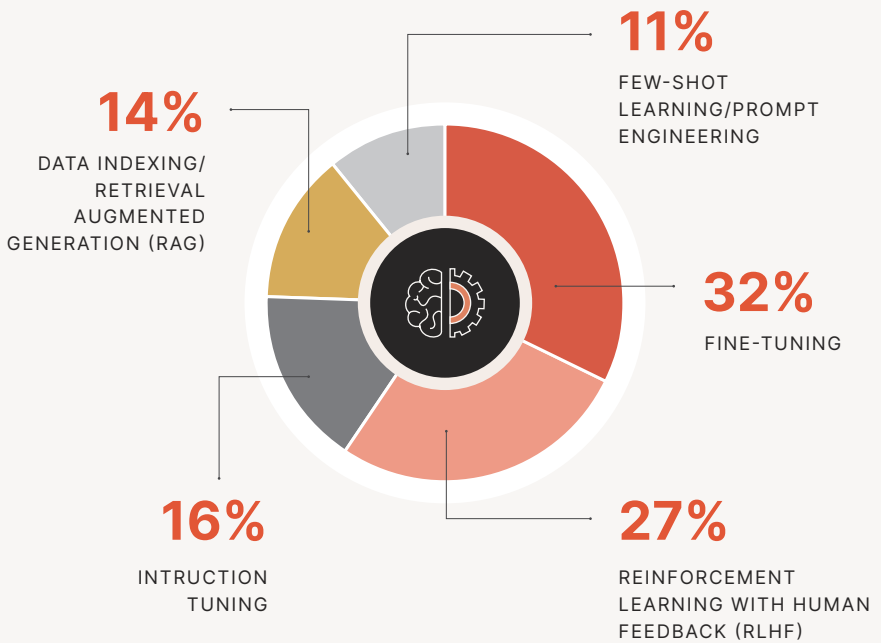
LLMs are instrumental in developing personalized recommenders, such as product recommendations based on customer preferences or purchase histories. Organizations can use LLMs to create content recommendations, such as movie suggestions on streaming platforms like Netflix, tailored to individual user interests and behaviors. In one application, a media company that we work with wants to let their customers ask questions like "What types of kid's movies featuring monsters would you recommend to me?" using previous movie history as a data point.

Customizing LLMs for Domain-Specific Tasks

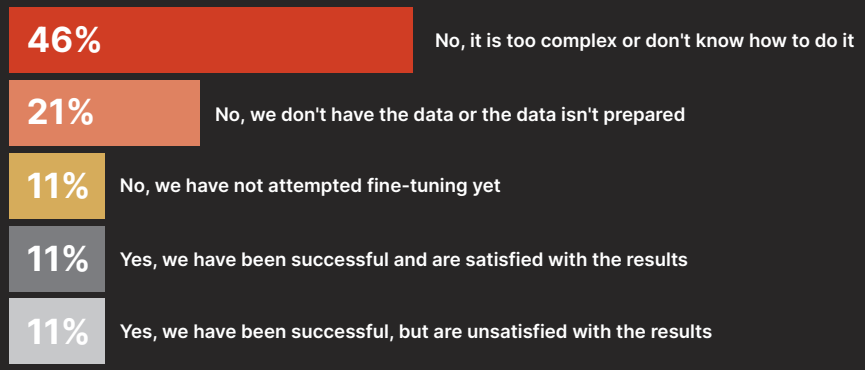
Our interviews made it clear that after initial experimentation and prompting, most teams want to customize their LLMs to achieve more accurate and tailored results. For example, it's been well documented that out-of-the-box LLMs struggle with code generation tasks when the LLMs aren't first fine-tuned. When asked how teams plan to approach customization, they overwhelmingly expressed an interest in fine-tuning and reinforcement learning with human feedback (RLHF) as their top two methods.

Fine-tuning is undoubtedly hot in the market. Every day there's a new social media post outlining a novel technique for fine-tuning LLMs, so we dug deeper and asked survey respondents about their level of success. Despite the interest, nearly half of respondents shared that fine-tuning is too complex. Another 21% said they don't have the data or still need to prepare it. Only 11% have successfully fine-tuned an LLM with optimal results, which illustrates a gap in the market for tools that make fine-tuning easy and approachable.

What technique are you most interested in using for LLM customization?



Have you successfully fine-tuned an LLM?

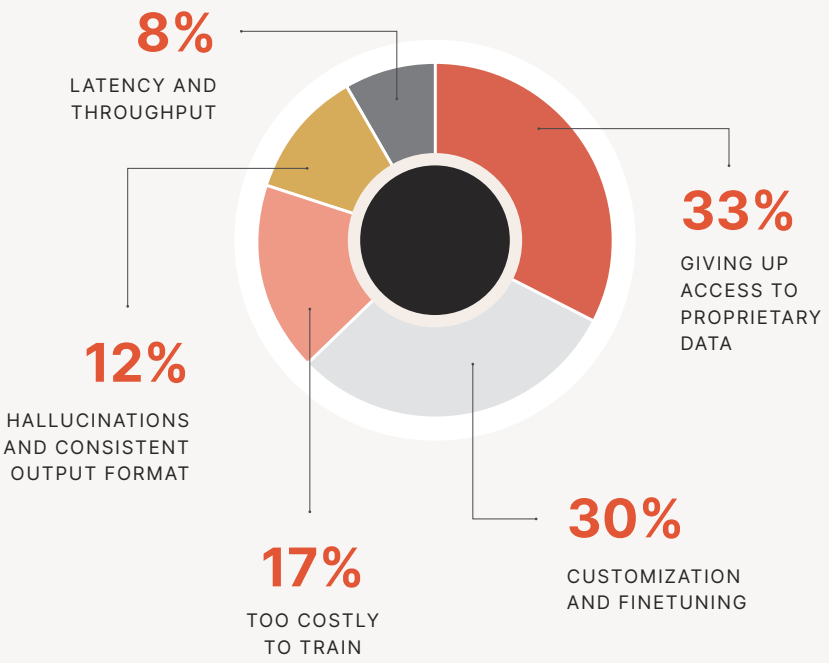


Do you have the requisite data for fine-tuning an LLM?

Key challenges when using LLMs in production

Throughout the survey, we captured both opportunities and challenges teams face when working with LLMs. Still, we wanted to get a clear picture of the top challenge when considering all the potential hurdles. As represented in the early questions, giving up access to proprietary data to commercial vendors and customization are the top challenges teams face when putting LLMs in production. The cost of training LLMs is also a growing concern, given their resource-intensive compute needs.

What is your top challenge preventing you from using LLMs in production?



> The biggest hurdle for LLMs right now is context. Today's LLMs do a great job of answering generic questions based on public data. That's **90%** of the journey, but that last mile is where they're struggling to provide contextual insights that are relevant to a business.

DEVVRET RISHI
Chief Product Officer, Predibase

Key challenges when using LLMs in production

EXPLORING LLM CHALLENGES AND SOLUTIONS

CHALLENGES	WHY IS THIS A CHALLENGE FOR LLMS?	HOW CAN YOU ADDRESS IT?
Latency	Several LLM applications like Chatbots, require low-latency sub-second response times, or you risk a poor user experience. Deploying LLMs on your infrastructure and optimizing for latency and compute costs requires deep expertise in infrastructure.	If you wish to avoid working with a commercial LLM vendor as it is oftentimes cost-prohibitive, consider working with an LLM infrastructure provider like Predibase that makes it easy to deploy and serve open-source LLMs within your cloud environment. By outsourcing the complexities of infra management, teams can focus on more valuable tasks like model customization.
Privacy	If you plan to rely on commercial vendors for your LLMs, then you need to get comfortable sharing sensitive data (e.g., customer data, IP, etc) over the internet via API calls. This is not an option for many organizations and regulatory requirements.	Deploy an open-source LLM within your virtual private cloud where your data remains secured, under your control. As an added benefit, by hosting the LLM in your environment, the model IP belongs to you as you explore customization.

CHALLENGES	WHY IS THIS A CHALLENGE FOR LLMS?	HOW CAN YOU ADDRESS IT?
Fine-tuning	Fine-tuning LLMs efficiently is complex, requiring both significant ML expertise and the ability to set up scalable and reliable training infrastructure.	Configuration-driven approaches like those provided by open-source Ludwig.ai make it easy to change model parameters in just a few lines of a configuration code. This makes it easy to iterate on fine-tuning techniques rapidly.
Costly to Train	In an ideal world, everyone would build their own LLM from scratch, training it on their own data, but this requires massive compute resources and is frankly out of reach for all but a handful of organizations.	Don't reinvent the wheel. Build on top of state-of-the-art open-source LLMs like LLaMa-2. Experiment with customization techniques such as fine-tuning and RAG to obtain similar results at a lower cost than training from scratch. Compression techniques can also be used to shrink LLMs for task-oriented jobs, further reducing inference costs.
Hallucinations	LLMs are known to occasionally provide responses that appear accurate but are not based on factual information.	Consider using techniques like Retrieval-Augmented Generation (RAG) to provide LLMs with factual information and reduce the likelihood of hallucinations. For traditional classification, don't forget that Supervised ML models may work just as well for your tasks. You can also employ fine-tuning to teach the model not to respond when it doesn't have sufficient information.

Key challenges when using LLMs in production

To understand how teams tackle these challenges with commercial solutions, we asked survey respondents what type of offferings they seek in the market today. Fine-tuning is clearly the #1 area of interest, but it's interesting to note that many organizations don't yet know what types of solutions they need.

What type of LLM solutions are you looking for in the market?

33%
PLATFORM TO CUSTOMIZE OR FINE-TUNE AN LLM

29%
DON'T KNOW YET WHAT WE NEED

21%
PLATFORM TO HOST AN LLM ON OUR PRIVATE DATA

17%
PLATFORM TO TRAIN AN LLM FROM SCRATCH

> We see a massive opportunity for customized open-source LLMs to help our teams generate real-time insights across our large corpus of project reports. The insights generated by this effort have big potential to improve the outcomes of our conservation efforts. We're excited to partner with Predibase on this initiative.

DAVE THAU
Global Data and Technology Lead Scientist, WWF

The path forward: Deploying customized open-source LLMs on a modern data stack

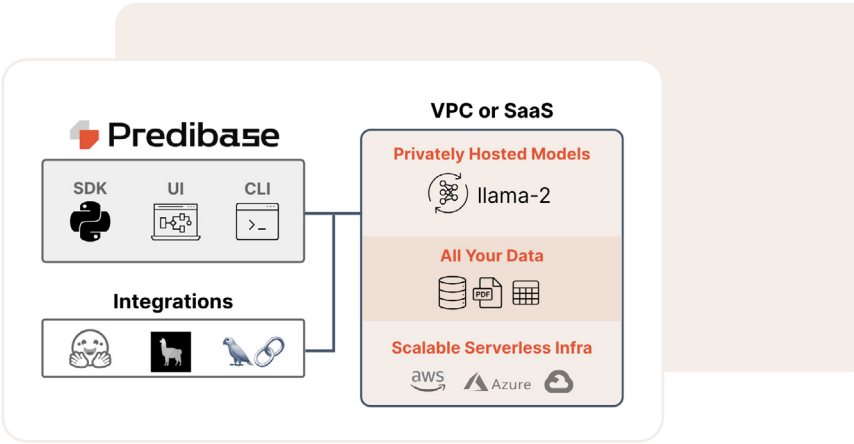
Based on our survey, it's clear: teams are looking for ways to customize and deploy open-source LLMs without giving up access to their proprietary data to a commercial vendor. Predibase addresses these challenges with the first LLM platform designed to help developers build AI-powered applications with custom open-source LLMs. Built on best-in-class managed infrastructure, Predibase provides the fastest way for your teams to deploy, operationalize and customize open-source LLMs on your data in your cloud.

> The declarative approach to ML makes it easy for our team of data scientists and engineers to train state-of-the-art language models on top of large volumes of unstructured data. With these insights, we're able to build more personalized user experiences.

HARSH SINGHAL,
Head of Machine Learning & AI, Koo Social

THE PREDIBASE ADVANTAGE

Predibase is the quickest and easiest way to build custom LLM applications without relying on cost-prohibitive commercial LLMs.



Privately Host LLMs Out-of-the-Box

Deploy and prompt state-of-the-art open-source LLMs — such as LLaMa 2, Falcon 7B, and Vicuna — instantly within your VPC or on the Predibase cloud with scalable, serverless infrastructure. No more managing complex distributed architectures or dealing with OOM errors. Simply choose your compute engine and Predibase automatically scales compute for the demands of your job, so you only pay for what you need.

The path forward: Deploying customized open-source LLMs on a modern data stack

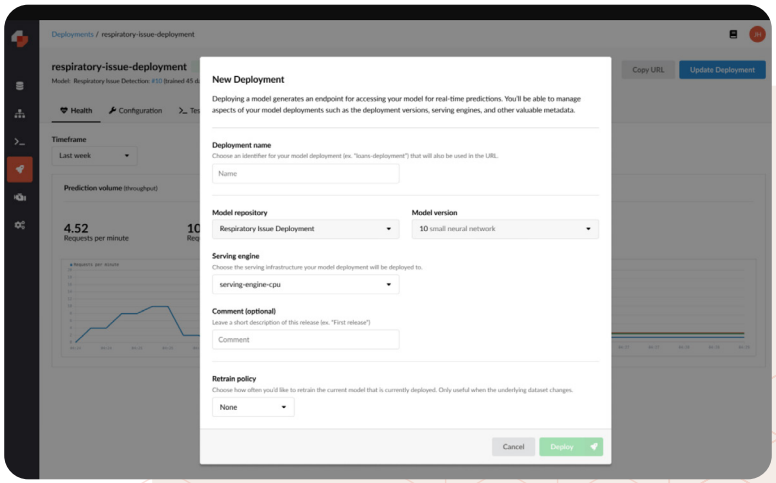
```python
import predibase as pb

prompt = "classify sentiment: this movie is amazing!

llm = pb.LLM("meta-llama/Llama-2-7b-hf").deploy()
response = llm.prompt(prompt)

finetuned_llm = llm.finetune(
    template="classify sentiment: {review}",
    target="sentiment",
    dataset="s3://reviews.csv",
).result()
response_ft = llm2.prompt(prompt)
```

Customize LLMs with Your Data

Stop relying on commercial LLMs — build on top of best-of-breed open-source models with Predibase. Integrations with popular open-source tools like LlamaIndex and HuggingFace enables you to seamlessly inject data into your prompts and leverage popular open-source models. Ready to fine-tune? Predibase offers a host of parameter-efficient techniques like LoRA supported by reliable scalable infrastrucuture. Predibase's declarative interface makes it easy to customize what you want while the platform automates the rest.

Operationalize LLMs with Ease

Build production-grade LLM applications on Predibase with serving options for real-time or batch inference and the ability to compress LLMs to improve performance and reduce costs. Prompt comparison tools and a robust set of model dashboards help you evaluate model performance, track changes over time and make smarter decisions.



Predibase

Built by AI leaders from Uber, Google, Apple, and Amazon and developed and deployed with the world's leading organizations.

payscale Uber Koo Adobe KOBLE WWF PARADIGM OPPLANE

TRY PREDIBASE FOR FREE