CASE STUDY: SCRIBD, INC.

Delivering faster, cheaper and more accurate results

Llama 3.1 8B, 70B and 405B power an AI content discovery assistant for Scribd, Inc.

Meta SCRIBD

At a glance

Scribd, Inc. used Llama 3.1 to deliver unmatched accuracy and cost savings for Ask AI, an AI content discovery assistant that generates personalized title recommendations and helps customers find the perfect read for the moment.

Industry
Technology/software

Use case
Enhancing real-time content discovery

Goal
Enabling faster, more personalized and engaging content discovery

Llama version
Llama 3.1 8B, 70B and 405B

Deployment
Amazon Web Services (AWS) and Databricks batch inference

Results*

76%
faster throughput (tokens per second) vs. previous model

97.7%
accurate macro-F1 score on intent detection

33%
compute cost savings with JSON output

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

The challenge

Limited search functionality, generalized recommendations

Scribd, Inc.'s Everand digital library subscription service made millions of bestselling and up-and-coming titles accessible to its users. But finding a title or topic relied primarily on keyword search, which had limited semantic and contextual understanding. Customers could browse recommendations, but the suggestions were based on pre-generated topics that they couldn't change.

These experiences made it difficult for users to find specific content and discover new titles that might interest them. To improve content discovery, Scribd, Inc. developed an AI-powered assistant called Ask AI.

Enabling faster, more personalized and engaging content discovery

For the newest version of Ask AI, Scribd, Inc. envisioned a generative AI assistant that could combine complete knowledge of Everand's complex catalog with a nuanced understanding of each customer to deliver amazing discovery experiences.

With a user base of 200 million unique monthly visitors and over 195 million pieces of content, the solution needed to meet a multi-layered set of AI requirements.

- Ensure accurate results with rich metadata for the full Everand catalog.
- Serve live customers effectively with a lean, fast model augmented with user behavioral and catalog data.
- Improve accuracy, optimize performance and manage infrastructure costs by fine-tuning models.

SCRIBD

Scribd, Inc. is a multinational tech company focused on the written and spoken word. The company's three products — Everand, Scribd and SlideShare — deliver knowledge, information and inspiration to billions of people across the globe.

Everand is home to millions of ebooks, audiobooks, podcasts and magazines. This global library includes bestselling and iconic titles, alongside diverse voices and emerging talent.

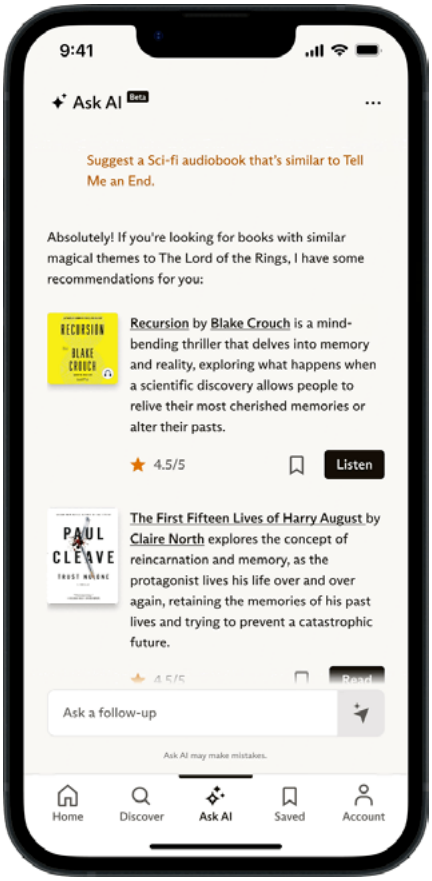
- **Industry:** Technology/software
- **Company size:** 300+ employees, 200 million monthly unique visitors

The solution

An AI content discovery assistant powered by Llama

Llama models immediately stood out as strong contenders. As an open-source product, Llama offered a cost-effective solution that was highly adaptable. When compared with other closed and open-source models, Llama delivered the essential component for success: high accuracy. Llama's unmatched accuracy and semantic understanding provided the nuanced responses Scribd, Inc. needed to transform Everand's Ask AI service.

> "The open-source nature of Llama enabled rapid experimentation and customization to meet our unique user needs. It allowed us to stay at the forefront of innovation, adapting the model to improve content recommendations quickly. For companies like ours, open-source models provide flexibility, allowing us to innovate faster."
>
> Prabdheep Cheema
> Senior Machine Learning Engineer, Scribd, Inc.

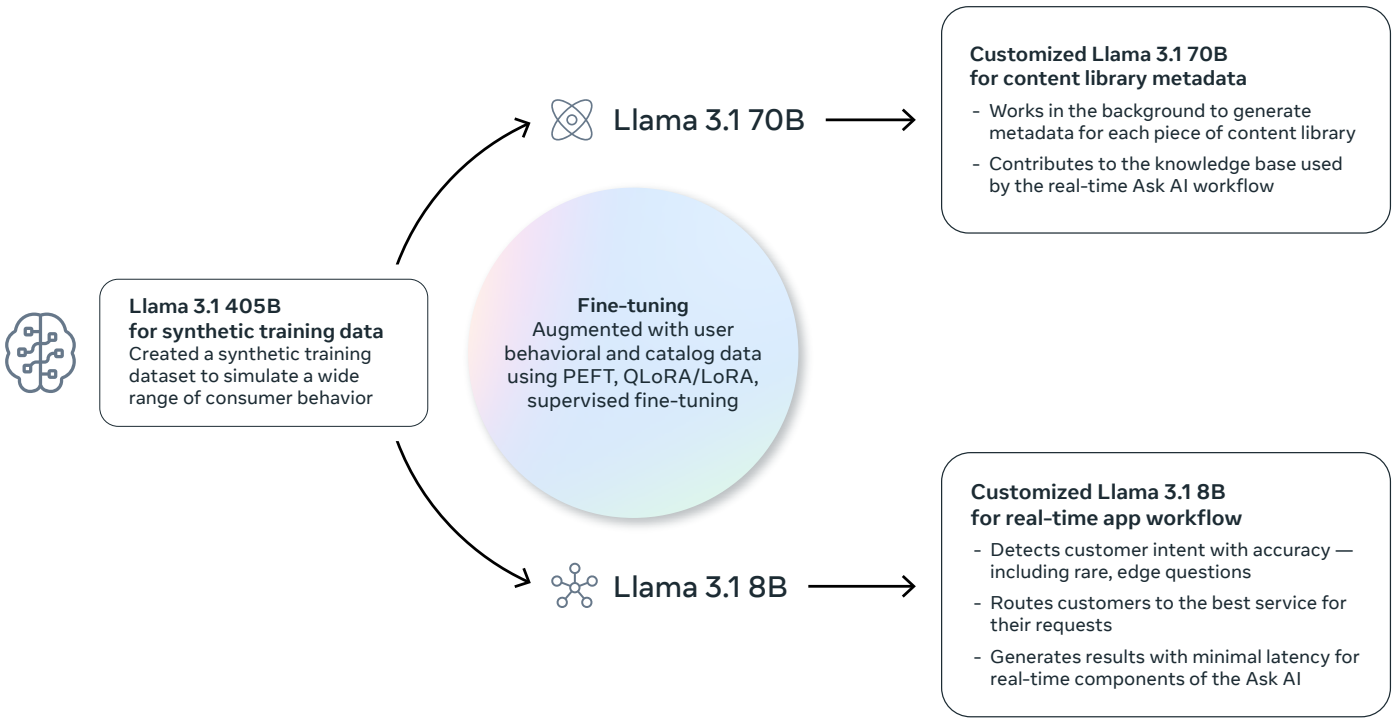


The new Ask AI experience provides on-point responses to nuanced questions.

A multi-layered, model distillation strategy

The Scribd, Inc. team used three Llama models to create the new Ask AI: Llama 3.1 405B, 8B and 70B. In addition to engineering an intuitive, knowledgeable AI agent, the team also worked to optimize performance and manage costs for every workload.

Everand customers explore an immense range of topics and ask practically every type of question, including obscure and unexpected ones, such as "How do ancient martial arts techniques feature in modern romance stories?" The team used Llama 3.1 405B to create synthetic data for a training dataset that simulated a wide range of consumer behavior. They then used parameter-efficient fine-tuning (PEFT) with QLoRA/LoRA and supervised fine-tuning to create a highly accurate, tailored version of Llama 3.1 8B.



Ask AI uses multiple Llama models to achieve performance.

Thanks to Llama's open-source nature, Scribd, Inc. was able to push beyond closed model limitations and achieve deeper customization. The retrained model was able to accurately detect customer intent — including the intent of rare, edge questions — then route customers to the best service for their requests.

By fine-tuning Llama 3.1 8B, the team was able to deliver better results with minimal latency for real-time components of the Ask AI app while managing the model's footprint and computing demands.

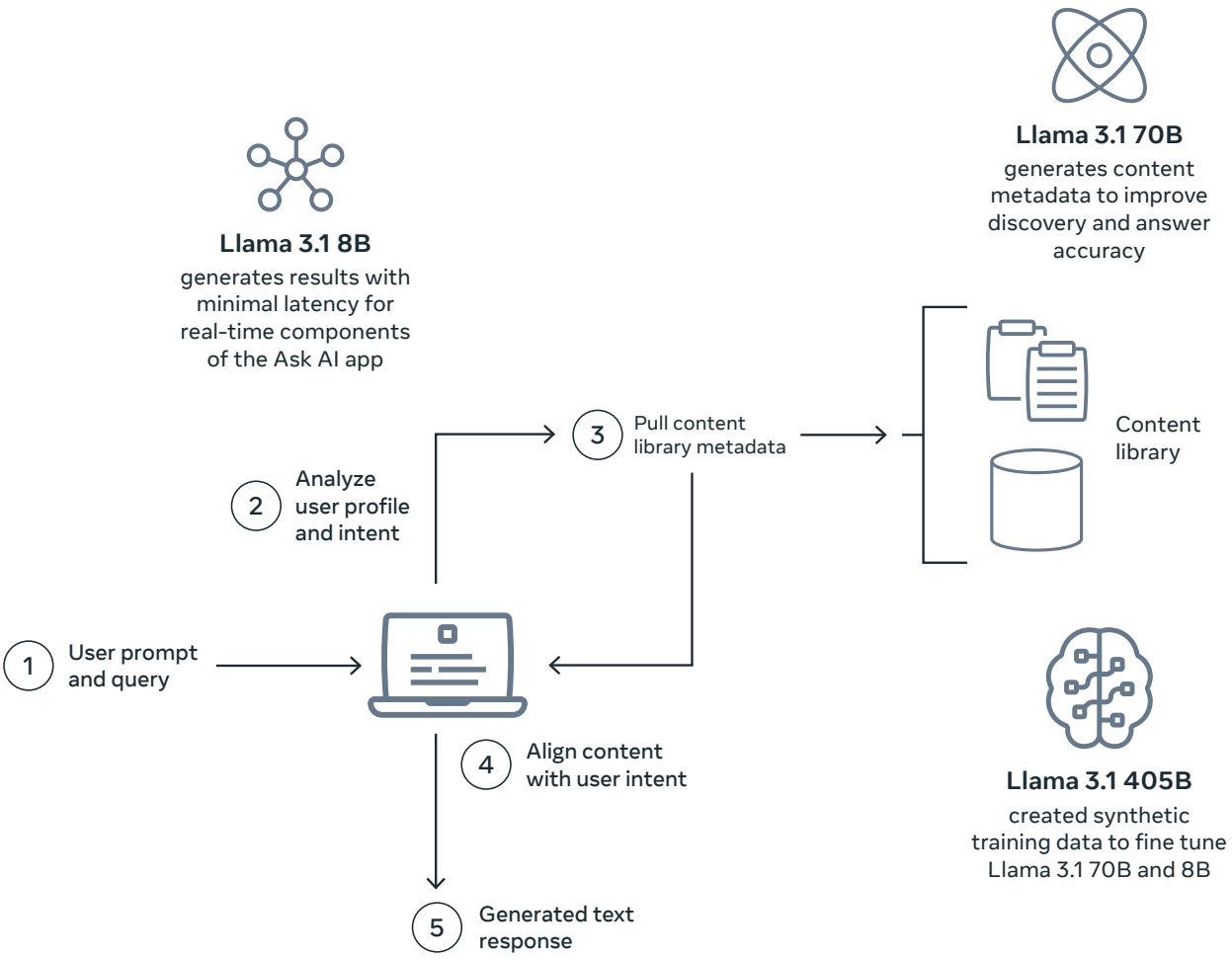
Llama 3.1 70B works in the background to generate metadata for each piece of content in the Everand library. This metadata contributes to the knowledge base used by the real-time Ask AI workflow, improving content discovery and answer accuracy.

> "The most important factor was creating a magical content discovery experience. Llama was our choice because of its superior ability to understand user intent and deliver accurate results quickly."
>
> Steve Neola
> Senior Director of Product, Generative AI, Scribd, Inc.

Getting to market fast with a compatible, customizable solution

Thanks to Llama offering the most comprehensive deployment options of any model provider, the team quickly found an approach compatible with its existing systems. Scribd, Inc. was able to integrate the model into its Ask AI assistant workflow without any major infrastructure changes.

The company used AWS along with Databricks batch inference to ingest and analyze massive amounts of data and support solution development. The application delivers the model's structured output in JSON format to improve metadata extraction and ensure high-quality, real-time responses.



The Ask AI application workflow is built to tailor recommendations to the user.

The outcome

A streamlined content discovery process strengthens customer relationships

The new Ask AI takes discovery far beyond discrete, specific title searches. The retrained Llama 3.1 8B model at the heart of the service has a nuanced understanding of customer intent and the Everand library that allows it to streamline the content discovery process by generating intuitive recommendations based on plot types, settings, genres and even other books a user likes.

Llama 3.1 has helped Everand deliver a better, more personalized customer experience that sets the brand apart and strengthens customer relationships.

At the same time, Llama has optimized the application for Scribd, Inc. In fact, implementing Llama resulted in major speed, cost and accuracy advantages compared with Scribd, Inc.'s original production model for Ask AI.

Using Llama 3.1 8B, the company saw 76% faster throughput in tokens per second and 33% compute cost savings with JSON output. That only augments the cost efficiencies already gained from selecting an open-source model. When it comes to accurate intent detection, the Llama-based application achieved an impressive 97.7% macro-F1 score.

76%
faster throughput in tokens per second

97.7%
accurate macro-F1 score on intent detection

33%
compute cost savings with JSON output

> "Llama stood out because of its superior performance in understanding broad user intents, which is critical for content discovery. It offered a cost-effective solution with high accuracy and speed, outperforming other options."

Steve Neola
Senior Director of Product, Generative AI, Scribd, Inc.

Conclusion

A commitment to reimagining content discovery with AI

Over time, Ask AI will mature into a next-level discovery agent with the power to enhance customer retention, improve loyalty and create higher customer lifetime value. Ask AI is just the beginning of their aspirations for Llama. Scribd, Inc. plans to integrate Llama into more areas of its user experience and tap into Llama Guard 3 for additional content trust and moderation support.

How can Llama enable your business?

Discover more about developing generative AI applications with open-source Llama large language models that bring unmatched control, customization and flexibility.

Learn More

Related Stories ▸

Meta