CASE STUDY: INFO EDGE

Empowering students to conduct better, faster research on higher education

Llama-enabled ShikshaGPT acts as an advisor for considering potential colleges and courses

Meta | infoedge

At a glance

Shiksha.com, part of Info Edge, set out to make academic research faster and more personalized for millions of students. The team used the compact, cost-effective Llama 3.1 8B model to build a conversational chatbot, ShikshaGPT. The model delivers fast and accurate inference without straining compute resources, allowing the company to run the chatbot efficiently on-premises. Students can now quickly and easily get responses to questions about colleges and courses to make more informed decisions about higher education.

Industry
Technology

Use case
Enabling chat-based college and course research for students

Goal
Transform decision-making about higher education into a faster, more tailored experience

Llama version
Llama 3.1 8B

Deployment
On-premises

Results*

85%
net positive user feedback on the quality of ShikshaGPT answers

10%
of the mobile app's responses earned "intent to apply" actions from users

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

The challenge

Researching colleges was too complex and time-consuming

Students researching higher education in India and abroad often struggled to find reliable information about colleges, courses and exams. Navigating traditional portals and forums — which scattered information across multiple sources — took ample time and effort. Many search tools lacked contextual understanding, leading to off-base results that didn't match the intent of the students' questions.

It was clear that students needed a better, faster way to gather the right information and make informed decisions about their education. Shiksha.com, part of Info Edge, was solving this problem for millions of students by offering a single site for credible college and course details. But as generative AI evolved, they saw an opportunity to take their solution to the next level.

Speeding up and personalizing college research

The Shiksha team wanted to use generative AI to make the college research process faster and more personalized. They needed a conversational chat solution for their site that could answer student questions accurately and in real time with targeted guidance, all while optimizing performance and costs. The large language model (LLM) behind the chat solution had to be powerful and fast enough to elevate their user experience, yet practical and efficient enough to scale as demand grew. That's where Llama came into play.

infoedge

Info Edge India Limited's online platform, Shiksha.com, empowers students to make informed education and career decisions. The platform provides them with credible, updated and personalized information about colleges, courses, exams and career paths in India and abroad.

- **Industry:** Technology
- **Company size:** 4000+ employees in 62 offices across 43 cities throughout India

"We wanted to provide students with a real-time, dynamic guidance engine to get quick advice in an easy-to-use, conversational format — so they could chart smarter academic paths with confidence."

Abhinav Katiyar,
Executive Vice President, Info Edge

The solution

A scalable and cost-efficient Llama-enabled chat solution for students

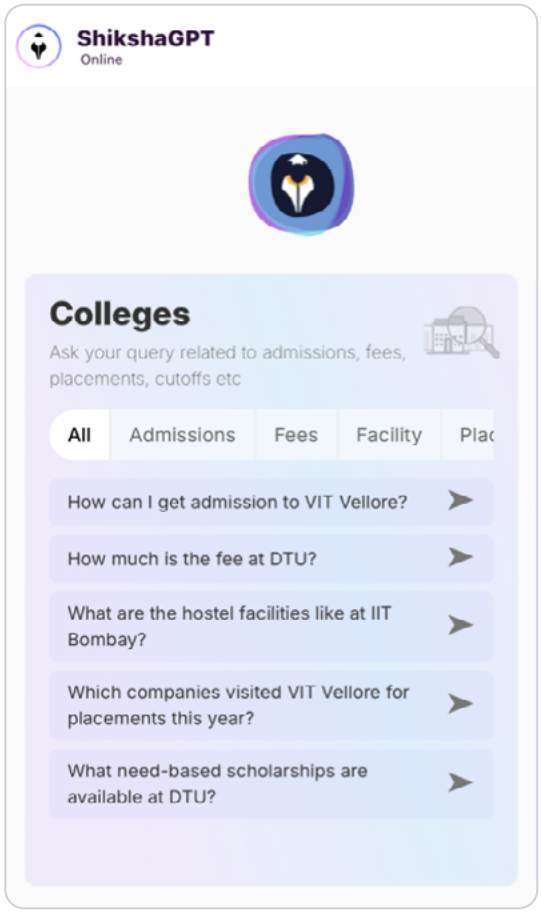
The team created an AI chatbot, ShikshaGPT, using Llama 3.1 8B. The chatbot, also available in WhatsApp, enables students to ask specific questions about colleges, courses and exams and get personalized, conversational answers fast.

After evaluating multiple LLMs, the team chose the compact Llama 3.1 8B model for ShikshaGPT for its cost advantages and lower computational requirements. While larger closed options could offer strong accuracy, they came with significant tradeoffs. Those tradeoffs included high compute demands and rising per-token inference fees that made scaling financially unsustainable.

In contrast, Llama 3.18B struck the right balance of cost and performance, and made it feasible to deploy ShikshaGPT on Shiksha's on-premise GPU infrastructure. The team could maintain total control over user data, system prompts and content and keep data secure. Ultimately, the model could run more efficiently in Shiksha's resource-constrained, on-premise environment, making the AI chatbot architecture more scalable. With fast inference times, the model served multiple users concurrently without performance issues or delays, improving the user experience.

During testing, Llama 3.1 8B demonstrated proficiency with academic language and aligned responses closely with student intent, ensuring that students received on-point information. Plus, its open-source nature provided flexibility and transparency, which helped the team deploy and iterate quickly.

> "The Llama 3.1 8B model was our perfect fit. It met our needs for response quality and fast inference with lower costs and compute requirements, allowing us to scale efficiently."
>
> **Romil Goel,**
> Vice President of Engineering,
> Info Edge



The intuitive ShikshaGPT enables students to research colleges and courses quickly and easily.

Semantic search meets structured content

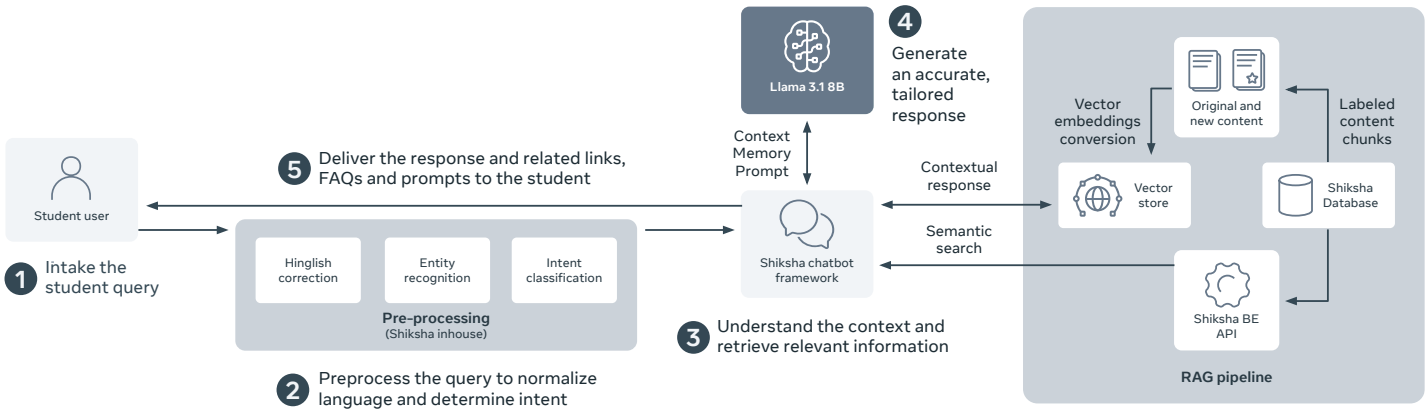
ShikshaGPT uses Llama 3.1 8B and a retrieval-augmented generation (RAG) workflow to understand complex educational queries and provide accurate responses based on information from curated academic databases, syllabi, university documents and more. The solution can also refine language and grammar for user-submitted questions and classify user intent for better responses.

Let's look closer at how the end-to-end ShikshaGPT architecture works. First, the student query flows through three pre-processing layers:

1. **Hinglish correction:** A spelling + Hinglish correction layer to standardize language in both English and Hindi
2. **Entity recognition:** An entity recognition layer to extract relevant academic terms, such as course names or universities
3. **Intent classification:** An intent classifier to understand the purpose of the query

These inputs are then routed into the Shiksha chatbot framework, which manages the conversation workflow, constructs the query prompt and prepares context for the Llama model. To enrich the context, the chatbot framework performs a semantic search using a vector store that contains embeddings of curated content from the Shiksha database. This includes both original and continuously updated academic material. Relevant content chunks are fetched through the Shiksha BE API, converted into embeddings and matched to the query before being passed to Llama.

With these structured inputs, memory and user intent, the Llama model generates a tailored, human-like response. That response is then returned to the student user along with related links, FAQs and follow-up suggestions.



The Llama-enabled ShikshaGPT architecture combines in-house natural-language processing layers and vectorized academic content to deliver accurate, personalized responses.

The outcome

Supporting faster and better college research at scale

Integrating Llama 3.1 8B into ShikshaGPT has empowered the site to deliver more personalized and engaging student research experiences. And the students they serve are impressed. In fact, Shiksha.com saw 85% net positive user feedback on answers from the new chat solution. This feedback signals that students not only trust the responses but are using them to make meaningful academic decisions.

ShikshaGPT has also fostered revenue growth for the site. More than 10% of the chatbot's responses on its mobile app drove users to take specific actions that confirmed their intent to apply to a college or university, a conversion metric that is tied to revenue generation for Shiksha. Ultimately, the compact Llama 3.1 8B model helped take ShikshaGPT from a content retrieval tool to a dynamic AI-enabled education advisor that delivers value to both students and Shiksha's business. With Llama, the team is now able to scale the chat solution to serve more students at a fraction of traditional costs.

85%
net positive user feedback on the quality of ShikshaGPT

10%
of the mobile app's responses earned "intent to apply" actions from users

"With compact Llama 3.1 8B at the core of ShikshaGPT, we're delivering smarter and more personalized guidance to millions of students cost-efficiently and securely at scale."

Romil Goel,
Vice President of Engineering, Info Edge

Conclusion

Expanding AI-enabled educational guidance

Looking ahead, Shiksha.com plans to continue advancing ShikshaGPT and improving satisfaction scores by evaluating evolving Llama models such as Llama 4 Scout for enhanced Hinglish comprehension, entity recognition and intent classification. With a scalable and cost-effective foundation in place, the team is well-positioned to reach more students across India and abroad — and continue making academic research more personalized, intuitive and effective for students.

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