

Fine-tuned LLaMA

3.1 Instruct with PEFT LoRA for Indian BFSI Domain



The better the question.
The better the answer.
The better the world works.



Shape the future
with confidence

Large Language Models (LLMs) are transforming natural language processing, excelling in diverse areas from general understanding to specialized technical tasks. However, applying them in heavily regulated sectors like banking and Non-Banking Financial Companies (NBFCs) brings unique challenges, particularly in handling domain-specific financial queries, ensuring regulatory compliance, and maintaining factual accuracy. This finetuned LLM is designed to tackle the unique challenges faced by the Indian BFSI sector: it overcomes the limitations of current models present in the market by providing deeper insights into BFSI-Specific terms such as 'missed call banking', 'revolving', 'DPD', 'foreclosure' etc. and it understand the compliance documentation and circulars for Indian BFSI regulatory standards. In addition, the traditional, resource-intensive fine-tuning process for large models becomes unsustainable with frequent regulatory updates, especially in multilingual markets like India.

In our approach, which employs Parameter Efficient Fine-Tuning (PEFT) with LoRA (Low-Rank Adaptation) on LLaMA 3.1 8B Instruct, we create a banking-specific model that maintains general capabilities while adapting to regulatory and language needs. Trained on RBI circulars and industry-specific data, this model at present supports English and Hindi, ensuring accurate, compliant and professional responses.

To benchmark the model's effectiveness, we have created an evaluation framework using Sem Score, a semantic similarity metric that captures contextual accuracy more effectively than traditional metrics like ROUGE or BLEU. The PEFT-based approach provides a scalable, cost-effective solution for LLM adaptation, establishing a new standard in regulatory compliance and multilingual support, specifically tailored for the financial sector.

As efficient execution of fine-tuning and evaluation processes is crucial in shaping the model's performance, in collaboration with NVIDIA and Neysa, EY has put in place state-of-the-art infrastructure - two H-100 GPUs hosted on Neysa servers.

Current challenges in adapting LLMs for Indian banking

Despite the success of models like FinGPT and BloombergGPT in addressing financial domain-specific needs, there are significant challenges in tailoring these models for the Indian banking sector. While FinGPT focuses on stock market data for real-time applications, and BloombergGPT is a high-performing, 50-billion-parameter model trained on English financial documents, both lack capabilities essential to the Indian banking and finance ecosystem. Key challenges include:

- **Language limitations as architecture is English-centric:** Both models are primarily trained on English corpora, and therefore lack support for Indian languages such as Hindi. Another unique requirement in India is the ability to handle code-mixed queries (e.g., "Hinglish"), which is very common in Indian financial communications.
- **Limited Indian BFSI context**
 - **Lack of knowledge on Indian banking:** FinGPT, for instance, is predominantly stock market-focused and lacks specific banking knowledge, impacting its ability to address queries on loan eligibility, account details, and other banking needs. For example, a query like "Mujhe gold loan ke liye kaise apply karna hai?" (How do I apply for a gold loan?) might be inadequately processed, as the model lacks familiarity with gold loans, a vital credit source in rural India.

Questions on Kisan Credit Cards (KCC) or PM Awas Yojana (PMAY) schemes could also go unanswered due to insufficient exposure to such India-specific products.

- **Cultural and contextual gaps:** These models often struggle with local financial terminology, regional banking terminology, and India-specific banking inquiries, reducing their efficacy in rural or culturally distinct settings. For example, queries like "Chit funds ke rules kya hote hain?" (What are the rules for chit funds?) might be misinterpreted, as chit funds are unique to certain Indian regions.

Similarly, questions about Self-Help Group (SHG) loans could be mishandled due to the model's lack of knowledge of microfinance practices in rural India.

- **Regulatory constraints:** The models do not encompass India's regulatory standards, such as Reserve Bank of India (RBI) guidelines on priority sector lending, KYC requirements, or policies like Jan Dhan Yojana, which are crucial for the country's financial inclusivity goals.

The limitations mentioned suggest that an Indian banking LLM would need:

- Capabilities to support regional Indian languages
- Inclusion of RBI and India-specific banking regulations during model training

- Cultural adaptation to address local banking terminology and practices.

These adaptations are essential for creating a language model that can effectively cater to India's unique and diverse banking landscape. The EY tool addresses these challenges effectively by using the appropriate datasets, fine-tuning them, testing and running comprehensive analyses of the test results.

Datasets

- **Data collection and sources:** Our dataset development was aimed at creating a robust knowledge base tailored for the banking and financial services industry. Key data sources included:
 - **RBI circulars:** We used 70 recent documents from the RBI to capture up-to-date banking regulations.

Example Q&A

Question: How are frauds classified according to the RBI Master Circular?

Answer: Frauds are classified based on the provisions of the Indian Penal Code into misappropriation and criminal breach of trust, fraudulent encashment through forged instruments, unauthorized credit facilities, negligence and cash shortages, cheating and forgery, irregularities in foreign exchange transactions, and other types of frauds.

Example Hindi Q&A

प्रश्न: RBI मास्टर सर्कुलर के अनुसार धोखाधड़ी को कैसे वर्गीकृत किया जाता है?

उत्तर: धोखाधड़ी को Indian Penal Code के provisions के आधार पर गलत तरीके से property का उपयोग, criminal विश्वास का उल्लंघन, धोखाधड़ी के माध्यम से निकासी, बिना permission वाली credit सुविधाएं, लापरवाही और cash की कमी, धोखा और ठगी, foreign currency लेन-देन में irregularities, और अन्य प्रकार की धोखाधड़ी में वर्गीकृत किया जाता है।

Example of a sample RBI circular Q&A data, along with its converted Hindi/Hinglish version

- **Bank and NBFC websites:** FAQs and customer service content were collected from official bank and NBFC websites.

Example Q&A

Question: What all transactions will be eligible for EMI conversion?

Answer: The value of the credit card transaction should be Rs. 2,500 or more to be eligible for conversion of transaction to EMI. The transactions like Jewelry/Gold, cash withdrawals, Balance Transfer and card fees will not be eligible for this facility.

Example Hindi Q&A

प्रश्न: कौन-कौन से लेन-देन EMI conversion के लिए योग्य होंगे?

उत्तर: क्रेडिट कार्ड लेन-देन का amount 2,500 रुपये या उससे अधिक होना चाहिए ताकि उसे EMI में परिवर्तित किया जा सके। आभूषण/सोना, cash withdrawals, Balance Transfer और कार्ड शुल्क जैसी लेन-देन इस सुविधा के लिए योग्य नहीं होंगे।

Example of sample bank and NBFC FAQ data with its Hindi/Hinglish translation

- **Bitext database:** A bilingual dataset containing customer support dialogues across retail banking, insurance, mortgage, and loan services.

Example Q&A

Question: I would like to activate a credit card online, will you help me?

Answer: I'm here to assist you with activating your credit card online. Activating your credit card is a simple and convenient process. Here's what you need to do:
 1. Visit the website of your credit card provider.
 2. Look for the "Activate Card" or "Card Activation" option. It is usually located in the "Services" or "Customer Support" section.
 3. Click on the option and follow the prompts to enter the required information, such as your card number, expiration date, and security code.
 4. Once you've entered the necessary details, verify the information and submit the activation request.

Example Hindi Q&A

प्रश्न: मुझे ऑनलाइन क्रेडिट कार्ड activate करना है, क्या आप मेरी मदद करेंगे?

उत्तर: मैं आपके क्रेडिट कार्ड को ऑनलाइन activate करने में मदद करने के लिए यहाँ हूँ। अपने क्रेडिट कार्ड को activate करना एक आसान प्रक्रिया है। आपको बस ये करना है:
 1. अपने क्रेडिट कार्ड provider की वेबसाइट पर जाएँ।
 2. Activate Card या "Card Activation" विकल्प खोजें। यह आमतौर पर "Services" या "Customer Support" सेक्शन में होता है।
 3. इस विकल्प पर क्लिक करें और ज़रूरी जानकारी डालने के लिए गए निर्देशों का पालन करें, जैसे कि आपका कार्ड नंबर, expiry date, और security code।
 4. एक बार जब आप सारी जानकारी भर लें, तो उसे चेक करें और activate करने का request सबमिट करें।

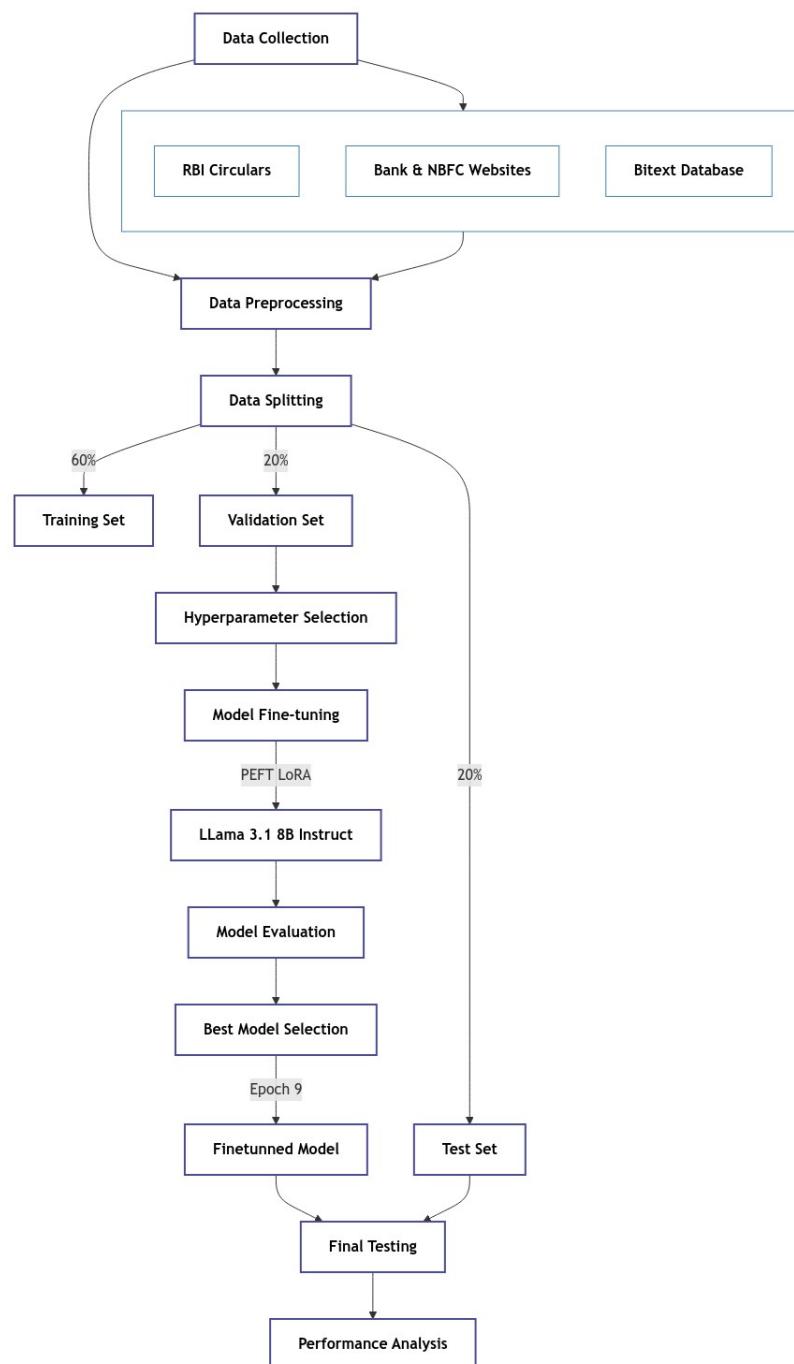
Example of sample bank and NBFC FAQ data with its Hindi/Hinglish translation

- **Data distribution and aggregation:** Data was meticulously categorized to cover various banking services, including retail loans, accounts, cards, and fund transfers. It can efficiently manage conversations across 29 BFSI-specific themes with 108 sub-categories in a remarkably better way with lower cost of infrastructure. We maintained distinct subcategories for services like ATM disputes, credit card activation, and loan applications to ensure comprehensive coverage.

RBI Guidelines	Loan QnA	Insurance QnA
Compliance, Regulation & Audit	CONTACT	PRODUCT_INFORMATION
Guidelines, Policies & Procedures	contact_agent	information_auto_insurance
Capital, Assets & Market Exposure	customer_service	information_health_insurance
Risk Management	human_agent	information_home_insurance
Customer Service & Account Management	FEES	information_life_insurance
Financial Reporting & Analysis	check_late_payment_fee	information_pet_insurance
Interest Rates & Pricing	lock_interest_rate	information_travel_insurance
Fraud Detection & Prevention	INFORMATION_REQUEST	CLAIMS
Retail Banking QnA		
ACCOUNT	borrowing_limit	accept_settlement
check_recent_transactions	check_fees	file_claim
close_account	check_loans	negotiate_settlement
create_account	compare_loans	receive_payment
ATM	estimate_loan_payment	reject_settlement
dispute_ATM_withdrawal	LOAN_APPLICATION	track_claim
recover_swallowed_card	apply_for_joint_loan	COMPLAINTS
CARD	apply_for_loan	appeal_denied_insurance_claim
activate_card	consolidate_debt	dispute_invoice
activate_card_international_usage	reapply_for_loan	file_complaint
block_card	LOAN_APPLICATION_PROCESS	CONTACT
cancel_card	change_application	agent
check_card_annual_fee	check_application_status	customer_service
check_current_balance_on_card	closing	insuranceRepresentative
CONTACT	submit_documentation	COVERAGE
customer_service	withdraw_application	change_coverage
human_agent	LOAN_MODIFICATIONS	check_coverage
FEES	add_coborrower	downgrade_coverage
check_fees	change_due_date	upgrade_coverage
FIND	extend_loan	ENROLLMENT
find_ATM	PAYMENT	buy_insurance_policy
find_branch	check_loan_terms	cancel_insurance_policy
LOAN	check_repayment_methods	cancellation_fees
apply_for_loan	make_additional_payments	compare_insurance_policies
apply_for_mortgage	pay_off_loan	INCIDENTS
cancel_loan	refinance_loan	report_incident
cancel_mortgage	request_payment_arrangement	schedule_appointment
check_loan_payments	split_payment	PAYMENT
check_mortgage_payments	turn_off_recurring_payments	check_payments
PASSWORD	turn_on_recurring_payments	payment_methods
get_password	PERSONAL_INFORMATION	report_payment_issue
set_up_password	change_personal_data	schedule_payments
TRANSFER	change_preferred_bank_account	POLICY
cancel_transfer	check_credit_report	change_personal_details
make_transfer	check_credit_score	QUOTE
	check_loan_details	calculate_insurance_quote
	check_privacy_policy	check_rates
		RENEW
		renew_insurance_policy

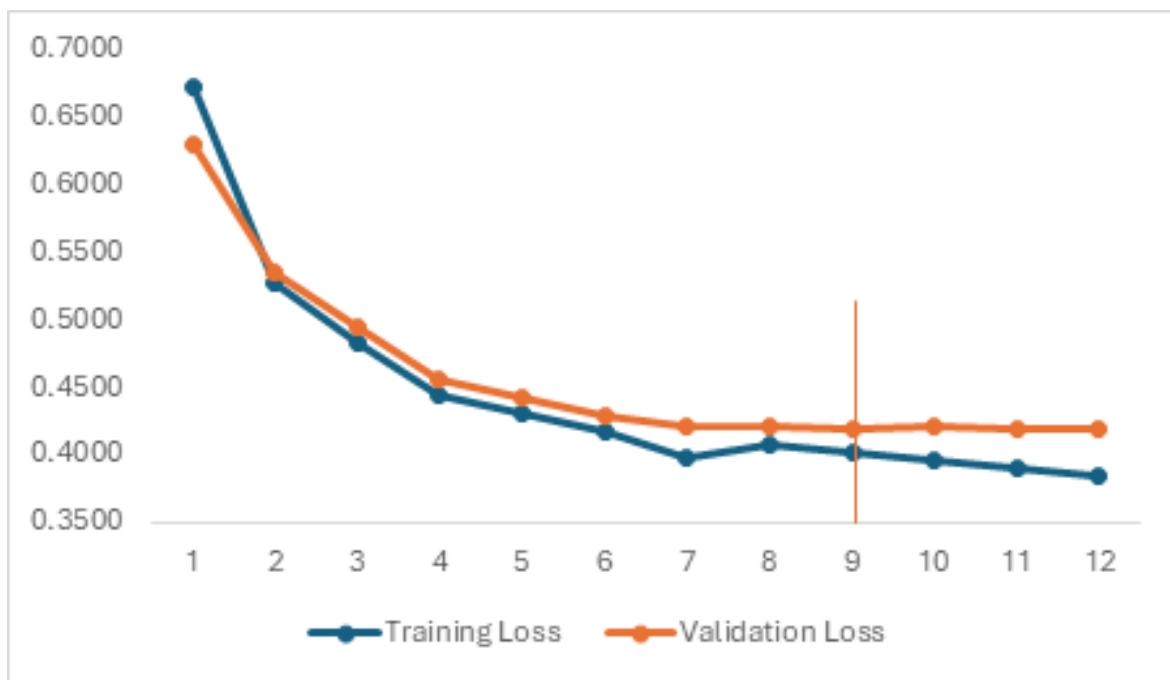
Data distribution across all collected banks and NBFCs FAQ data and RBI circular QnA data

- **Hindi dataset creation:** To make the dataset bilingual, we translated all English data into Hindi using the LLaMA 70B Instruct model, preserving specific banking terms like "RBI" and "credit" in English. This bilingual approach allows the model to serve diverse linguistic demographics while maintaining domain-specific terminology.
- **Approach from data to analysis:** The multi-source approach to data collection (RBI circulars, banks and NBFC websites, and Bitext databases) ensured comprehensive data coverage across regulatory and customer service topics, which is foundational for a robust financial language model.



Overall finetuning approach

- Data preprocessing and splitting: After cleaning and formatting the data into a structured question-answer format, we split the dataset into 60% training, 20% validation, and 20% testing sets. Splitting was performed category-wise (e.g., Account Management, ATM Services, etc.) to ensure each dataset had balanced category representation, critical for model generalization across varied queries.
- Model selection and fine-tuning: We chose the LLaMA 3.1 8B instruct model as our base, implementing Low-Rank Adaptation (LoRA) for parameter-efficient fine-tuning. This selection enabled domain-specific knowledge infusion while retaining the model's general capabilities, helping prevent "catastrophic forgetting" – a common challenge in fine-tuning large models on specific domains. (Read more here: [catastrophicforgetting paper](#))
- Hyperparameter optimization: We optimized LoRA parameters to maximize accuracy and efficiency. Key adjustments included:
 - Rank (r) values of 8, 16, and 32, selecting $r=32$ as it yielded optimal accuracy.
 - Alpha (α) values such as 16, 32, and 64, choosing $\alpha=64$ for balanced parameter scaling. These selections ensured the model adapted well to financial domain-specific language. (Read more here: [lora paper](#), [hyperparameteroptimization paper](#))
 - Training and overfitting prevention: Training was conducted using strategies to ensure stability and prevent overfitting:
 - Gradient clipping and weight decay helped control gradients and regularize the model. (Know more: [Gradientclipping paper](#))
 - Mixed Precision Training (FP16) optimized memory and speed, accelerating training without compromising performance. (Know more: [MixedPrecision paper](#))
 - Best Model Retention by automatically saving the model with the lowest validation loss. The validation loss plateaued at epoch 9, indicating it as the optimal stopping point to avoid overfitting.



- Model evaluation and final testing: Post-training, we evaluated the model on the test set across categories, using SemScores to assess alignment with ground truth answers. The final testing validated the model's readiness to handle domain-specific queries accurately, making it suitable for banking and NBFC applications. (Read more here: [semscore paper](#))

- **Evaluation framework**

- We have developed an evaluation framework that can be leveraged for any LLM, using SemScore to benchmark performance across various use cases. This framework enables us to objectively assess the accuracy and alignment of model outputs by focusing on semantic similarity, addressing the limitations of traditional methods like BLEU and ROUGE.
- In our recent benchmarking tests, we applied this framework to evaluate the fine-tuned LLaMA 3.1 model. The results demonstrate its clear superiority, outperforming both OpenAI and Sarvam's latest models across multiple tasks. These findings validate the effectiveness of our finetuned model in achieving superior results.

	LLAMA 3.1 - 8B Instrucrt		OpenAI		Sarvam AI		Finetuned Model
	Default	Finetuned	GPT 4o	GPT 4o mini	Sarvam2B	OpenHathi	
Banks & NBFC FAQs	53.04%	86.46%	63.70%	59.49%	50.23%	47.93%	
RBI Regulations	55.42%	87.95%	65.22%	61.35%	51.92%	49.22%	
Retail Banking	59.55%	94.35%	91.83%	77.16%	58.63%	52.32%	
Mortgage Loan	58.34%	94.21%	90.83%	72.85%	39.42%	46.35%	
Insurance	54.01%	98.20%	89.71%	68.37%	49.16%	43.36%	

What's next?

Building on the findings and contributions of this research, future work will focus on expanding the model's capabilities and addressing key challenges to enhance its practical applicability in the BFSI domain and beyond. Key areas of future exploration include:

1. **Development of Multi-Indic Language Capabilities:** Expanding the model's bilingual functionality to support additional Indian languages such as Marathi, Gujarati, Tamil, and Bengali, ensuring comprehensive financial inclusion across linguistically diverse regions. This will involve the creation of language-specific datasets and fine-tuning strategies to maintain semantic and contextual accuracy.
2. **Integration with Domain-Specific Systems:** Enhancing the model's functionality to enable seamless function calling and integration with essential banking software such as CRM, LMS, LOS, CMS, Card Management Systems, PMS, and Financial Data Lakes. The goal is to refine the model's ability to interact with these systems in a faster, more efficient, and scalable manner.
3. **Optimization of Inference Time:** Reducing inference time by employing advanced parameter pruning techniques tailored for BFSI-specific tasks. This optimization will focus on maintaining accuracy while significantly improving the model's performance for real-time applications.
4. **Evaluation of Quantized Model Versions for Edge Deployment:** Exploring quantized versions of the fine-tuned LLaMA 3.1 model to assess its viability for edge device usage. This will involve benchmarking the trade-off between reduced resource requirements and the model's semantic performance, ensuring robust performance in low-compute environments.
5. **Exploration of Advanced Evaluation Metrics:** Further refining the evaluation framework to include metrics that assess multi-lingual and multi-functional capabilities, particularly in regulatory and high-stakes scenarios. This will ensure the model's reliability and suitability for deployment in critical BFSI workflows.

Through these advancements, the research aims to create a more versatile, efficient, and inclusive LLM framework for regulated industries, contributing to both the academic and practical understanding of domain-specific AI applications.

NOTES

Our Offices

Ahmedabad

22nd Floor, B Wing, Privilon
Ambli BRT Road, Behind
Iskcon Temple
Off SG Highway
Ahmedabad - 380 059
Tel: + 91 79 6608 3800

8th Floor, Building No. 14A
Block 14, Zone 1
Brigade International
Financial Centre
GIFT City SEZGandhinagar
- 382355, Gujarat
Tel +91 79 6608 3800

Bengaluru

12th & 13th Floor
"UB City", Canberra Block
No.24 Vittal Mallya Road
Bengaluru - 560 001
Tel: + 91 80 6727 5000

Ground & 1st Floor
11, 'A' wing
Divyasree Chambers
Langford Town
Bengaluru - 560 025
Tel: + 91 80 6727 5000

3rd & 4th Floor
MARKSQUARE
#61, St. Mark's Road
Shantala Nagar
Bengaluru - 560 001
Tel: + 91 80 6727 5000

1st & 8th Floor, Tower A
Prestige Shantiniketan
Mahadevapura Post
Whitefield, Bengaluru - 560 048
Tel: + 91 80 6727 5000

Bhubaneswar
8th Floor, O-Hub, Tower A
Chandaka SEZ, Bhubaneswar
Odisha - 751024
Tel: + 91 674 274 4490

Chandigarh

Elante offices, Unit No. B-613 & 614
6th Floor, Plot No- 178-178A
Industrial & Business Park, Phase-I
Chandigarh - 160 002
Tel: + 91 172 6717800

Chennai

6th & 7th Floor, A Block,
Tidel Park, No.4, Rajiv Gandhi Salai
Taramani, Chennai - 600 113
Tel: + 91 44 6654 8100

Delhi NCR

Aikyam
Ground Floor
67, Institutional Area
Sector 44, Gurugram - 122 003
Haryana
Tel: + 91 124 443 4000

3rd & 6th Floor, Worldmark-1
IGI Airport Hospitality District
Aerocity, New Delhi - 110 037
Tel: + 91 11 4731 8000

4th & 5th Floor, Plot No 2B
Tower 2, Sector 126
Gautam Budh Nagar, U.P.
Noida - 201 304
Tel: + 91 120 671 7000

Hyderabad

THE SKYVIEW 10
18th Floor, "SOUTH LOBBY"
Survey No 83/1, Raidurgam
Hyderabad - 500 032
Tel: + 91 40 6736 2000

Jaipur

9th floor, Jewel of India
Horizon Tower, JLN Marg
Opp Jaipur Stock Exchange
Jaipur, Rajasthan - 302018

Kochi

9th Floor, ABAD Nucleus
NH-49, Maradu PO
Kochi - 682 304
Tel: + 91 484 433 4000

Kolkata

22 Camac Street
3rd Floor, Block 'C'
Kolkata - 700 016
Tel: +91 33 6615 3400

6th floor, Sector V, Building
Omega, Bengal Intelligent Park,
Salt Lake Electronics Complex,
Bidhan NagarKolkata - 700 091
Tel: +91 33 6615 3400

Mumbai

14th Floor, The Ruby
29 Senapati Bapat Marg
Dadar (W), Mumbai - 400 028
Tel: + 91 22 6192 0000

5th Floor, Block B-2
Nirlon Knowledge Park
Off. Western Express Highway
Goregaon (E)
Mumbai - 400 063
Tel: + 91 22 6192 0000

3rd Floor, Unit No 301
Building No. 1
Mindspace Airoli West (Gigaplex)
Located at Plot No. IT-5
MIDC Knowledge Corridor
Airoli (West)
Navi Mumbai - 400708
Tel: + 91 22 6192 0003

Altimus, 18th Floor Pandurang
Budhkar Marg Worli, Mumbai -
400 018 Tel: +91 22 6192 0503

Pune

C-401, 4th Floor
Panchshil Tech Park, Yerwada
(Near Don Bosco School)
Pune - 411 006
Tel: + 91 20 4912 6000

10th Floor, Smartworks
M-Agile, Pan Card Club Road
Baner, Pune - 411 045
Tel: + 91 20 4912 6800

Contact us



Rohit Pandharkar
Technology Consulting
Partner, EY India
rohit.pandharkar@in.ey.com



Durjoy Sen Maitra
Senior Manager,
EY India
durjoy.maitra@in.ey.com



Ernst & Young LLP

EY | Building a better working world

EY is building a better working world by creating new value for clients, people, society and the planet, while building trust in capital markets.

Enabled by data, AI and advanced technology, EY teams help clients shape the future with confidence and develop answers for the most pressing issues of today and tomorrow.

EY teams work across a full spectrum of services in assurance, consulting, tax, strategy and transactions. Fueled by sector insights, a globally connected, multi-disciplinary network and diverse ecosystem partners, EY teams can provide services in more than 150 countries and territories.

All in to shape the future with confidence.

EY refers to the global organization, and may refer to one or more, of the member firms of Ernst & Young Global Limited, each of which is a separate legal entity. Ernst & Young Global Limited, a UK company limited by guarantee, does not provide services to clients. Information about how EY collects and uses personal data and a description of the rights individuals have under data protection legislation are available via ey.com/privacy. EYG member firms do not practice law where prohibited by local laws. For more information about our organization, please visit ey.com.

Ernst & Young LLP is one of the Indian client serving member firms of EYGM Limited. For more information about our organization, please visit www.ey.com/en_in.

Ernst & Young LLP is a Limited Liability Partnership, registered under the Limited Liability Partnership Act, 2008 in India, having its registered office at Ground Floor, Plot No. 67, Institutional Area, Sector - 44, Gurugram - 122 003, Haryana, India.

EYIN2501-010

© 2025 Ernst & Young LLP. Published in India.
All Rights Reserved.

This publication contains information in summary form and is therefore intended for general guidance only. It is not intended to be a substitute for detailed research or the exercise of professional judgment. Neither EYGM Limited nor any other member of the global Ernst & Young organization can accept any responsibility for loss occasioned to any person acting or refraining from action as a result of any material in this publication. On any specific matter, reference should be made to the appropriate advisor.

SA1

ey.com/en_in

 @EY_India  EY  EY India  EY Careers India  @ey_indiacareers