ACM multimedia

Dublin, Ireland 27-31.10.2025

Sovereign & Shared: Frugally Scalable Multilingual-Multimodal AI (for Bharat?)

IIT BOMBAY

BharatGen

GenAI for Bharat, by Bharat

Dr. Maneesh Singh, VP, ML
(Bharatgen.com Consortium)

About Myself

Who am I?

Over 25 years in AI/ ML R&D

Developed and delivered AI models and systems for 20+ years in multiple sectors: Industrial AI, fintech (insurance, banking, energy sectors), security and surveillance, medical imaging and diagnostics, automatic driver-assistance and intelligent traffic systems.

('25-) VP of Machine Learning, BharatGen

Co-leading (one of) India's Sovereign AI efforts

- Joining the team on November 1, 2025

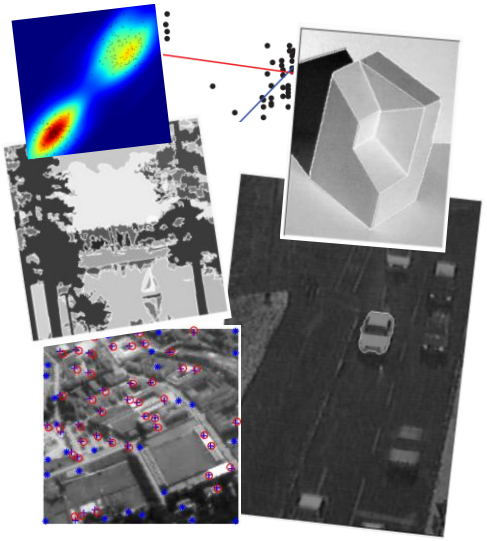
('16-'22) Head, AI R&D. Verisk Analytics

Co-created Verisk Innovative Analytics (Center of Excellence)

Other Engagements: President, Indic AI-Dias ('24-present). ('23-'25) Several Startups. Distinguished Research Scientist, Comcast AI & Discovery Lab ('23). Technology Leader, Vision & Leaning, SRI International ('13-'15),

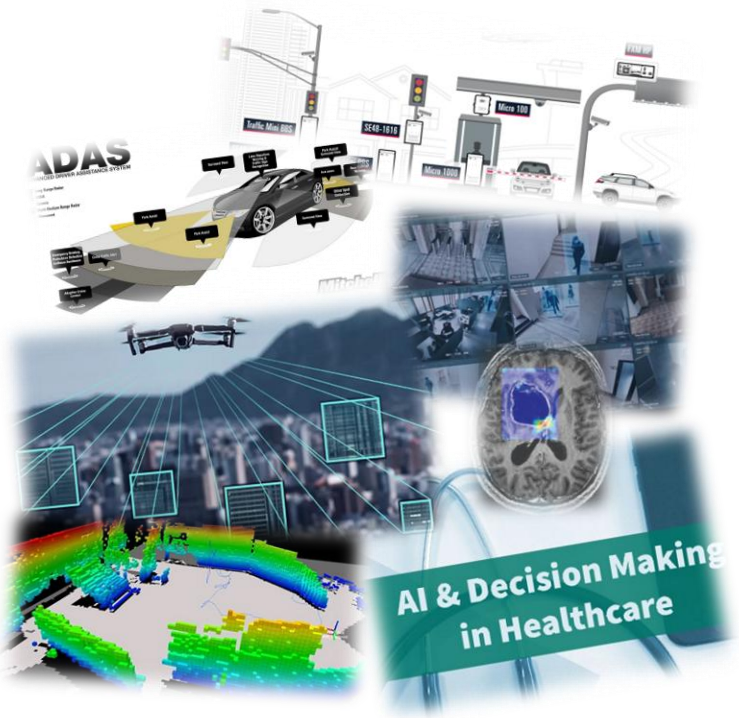
PhD ('03) in Computer Vision, UIUC (Advisor: Prof. Narendra Ahuja)

Major Stints



UIUC PhD ‘03 (Narendra Ahuja)

Nonparametric density estimation
Robust statistics, Segmentation, matching, registration, tracking…



Siemens, Staff Scientist (’04-’13)

Real-Time Vision Systems: ADAS, ITS, condition monitoring, Multicamera Surveillance, robotics and 3D SLAM, medical diagnostics and decision making …

Verisk, Head of AI R&D (’16-’22)

Digital media forensics, property intelligence, document understanding, weather forecasting and simulations, pandemic modeling, audio-based health prognostics, …

What is BharatGen?

A 5-minute YouTube Video on BharatGen

[Launching India’s Multimodal GenAI Future | BharatGen]

- https://www.youtube.com/watch?v=WPbZeFKNZm8

From a sidebar to a company

Feb 2023

- Chat with Prof. Ramakrishnan at AAAI ‘23

August 2024

- ‘Seed’ funding announced by DST, India

October 2024

- BharatGen created; first tranche received

January 2025

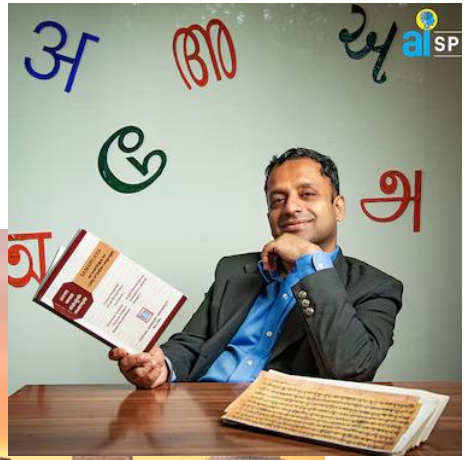
- Rishi Bal (ex- Google Research, Microsoft, CMU) took over as CEO

August 2024

- Major funding by India AI Mission announced

November 01, 2025

- Landing onsite as VP, Machine Learning



BharatGen: A National Generative AI Initiative for Inclusive Innovation

IIT BOMBAY

BharatGen

GenAI for Bharat, by Bharat

Rishi Bal, Executive Vice President, BharatGen
Dr. Maneesh Singh, VP, ML BharatGen

Pankaj Singh, Program Director, BharatGen

Prof. Ganesh Ramakrishnan
Indian Institute of Technology Bombay
Bank of Baroda Chair Professor in Digital Entrepreneurship,
CSE, IIT Bombay

PI: BharatGen Consortium

BharatGen Consortium

- Prof. Ganesh Ramakrishnan, IIT Bombay
- Prof. S. Umesh, IIT Madras
- Prof. Rohit Saluja, IIT Mandi
- Prof. Mohan Raghavan, IIT Hyderabad
- Prof. Ravi Kiran, IIIT Hyderabad
- Dr. Madhusudhanan ("Madhu") Baskaran, IIT Madras
- Prof. V. Kamakoti, IIT Madras
- Prof. Arnab Bhattacharya, IIT Kanpur
- Prof. Maunendra Desarkar, IIT Hyderabad
- Prof. Aditya Maheshwari, IIM Indore
- Bhashini, Digital India Bhashini Division, MeiTY

The BharatGen Consortium

BHASHINI

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

IIT Bombay

(PI) Prof Ganesh Ramakrishnan

IIT Madras

Prof V. Kamakoti
Prof S. Umesh
Prof Madhusudanan

BharatGen

(EVP) Rishi Bal
VP Maneesh Singh
VP Pankaj Singh

IIT Kanpur

Prof Arnab Bhattacharya

IIT Hyderabad

Prof. Maunendra Desarkar
Prof. Mohan Raghavan

IIIT Hyderabad

INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY
HYDERABAD

Prof. Ravi Kiran

IIT Mandi

Prof Rohit Saluja

IIM Indore

सिद्धिमूलं प्रबन्धनम्
भा. प्र. सं. इन्दौर
IIM INDORE

Prof Aditya Maheshwari

IIT Bombay

Prof Pratik Jawanpuria*

IIIT Delhi

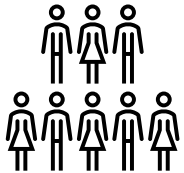
Prof Bapi Chatterjee*

IIT Kharagpur

Prof Pawan Goyal*

Other collaborators: Prof. Preethi Jyothi (IIT Bombay), Prof. Arpit Agarwal (IIT Bombay), Prof. Ranjith Padinhateeri (IIT Bombay), Prof. Ganesh Sivaraman (IIT Bombay), Prof. Nirmal Punjabi (IIT Bombay), Prof. Makarand Kulkarni (IIT Bombay), Prof. Jaideep Joshi (IIT Bombay), Prof. Varsha Apte (IIT Bombay)

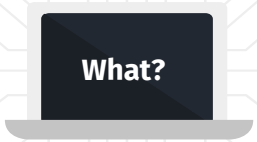
Scale: People and Funding



- **BharatGen**: 40 FTEs including leadership, management, scientsits and engineers
- **BharatGen Consortium**: 10+ institutes, 20+ faculty, and 50+ graduate students.
- Goal (2026): 100+ Full-time staff

- **Seed Funding**: Department of Science & Technology (DST), India announced funding of INR 235 cr (equiv. US $26.6 million) in August 2024.
- **India AI Mission:** announced additional funding INR 988.6 cr (equiv. US $112 million) in August 2025.
- Raising more money …

A Broad, Ambitious Mission



Technological sovereignty

Nurture an ecosystem for Industries, startups, innovators, research institutions to help achieve AI parity and independence for India

Data localisation

Bharat Data Sagar: Transform India into a global data powerhouse, by establishing India-centric benchmarks and leaderboards.

Digital Inclusion

Build solutions across domains ranging from agriculture, health education, defense to entertainment

Security & Privacy

Atma Nirbhar AI

Foster public private partnerships to achieve self-reliance in AI

Localised solutions

Foster AI inventions and turn India's diversity and cultural richness into an opportunity for India.

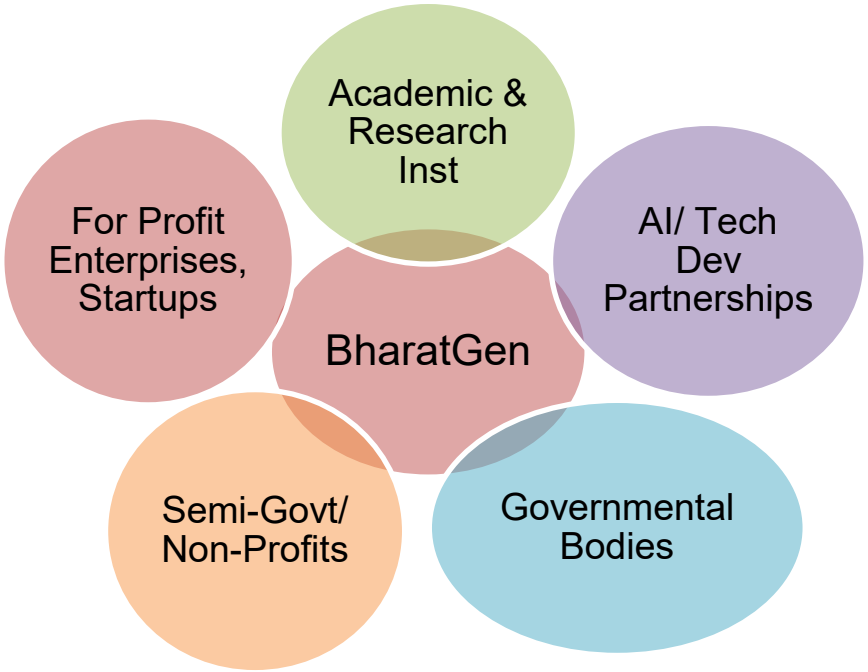
Cultural preservation

Generate high quality text, image, and speech modalities across various Indian languages and cultural contexts

Transition from IP consumer to IP producer

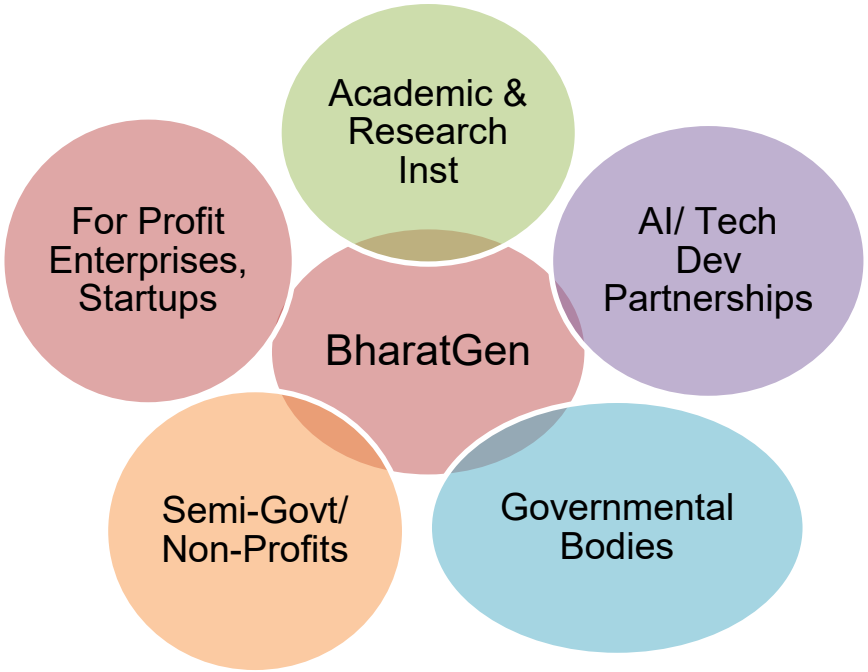
Product Centric Approach: Transition Bharat from IP Consumer to a large IP Generator in AI

Broad Partnership Profile



- **Academic and Research Institutions**
  - IITs, IIMs, IIITs; looking to expand globally
- **Governmental bodies: Central and State**
  - DARPG (Admin Reform; Public grievances), WASH (Water, Sanitation & Hygiene), ICAR (Indian Counsil of Agricultural Research), DGIS (Directorate General of Information Systems),
- **Semi-government / Non-profit organizations**
  - Eg: IBDIC (Indian Banks' Digital Infrastructure Company), IFSCA (Int'l Financial Services Centres Authority), Mass Entrepreneurship, Sansad TV, NABARD (National Bank for Agriculture and Rural Development), NPCI (National Payments Corporation of India), …

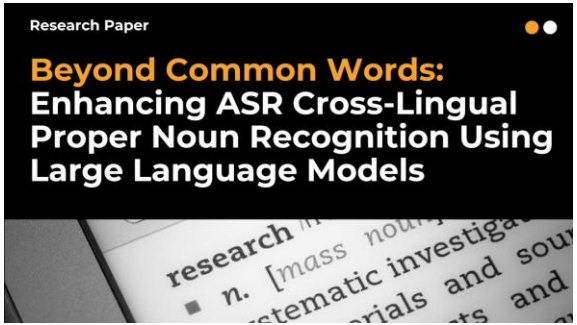
Broad Partnership Profile



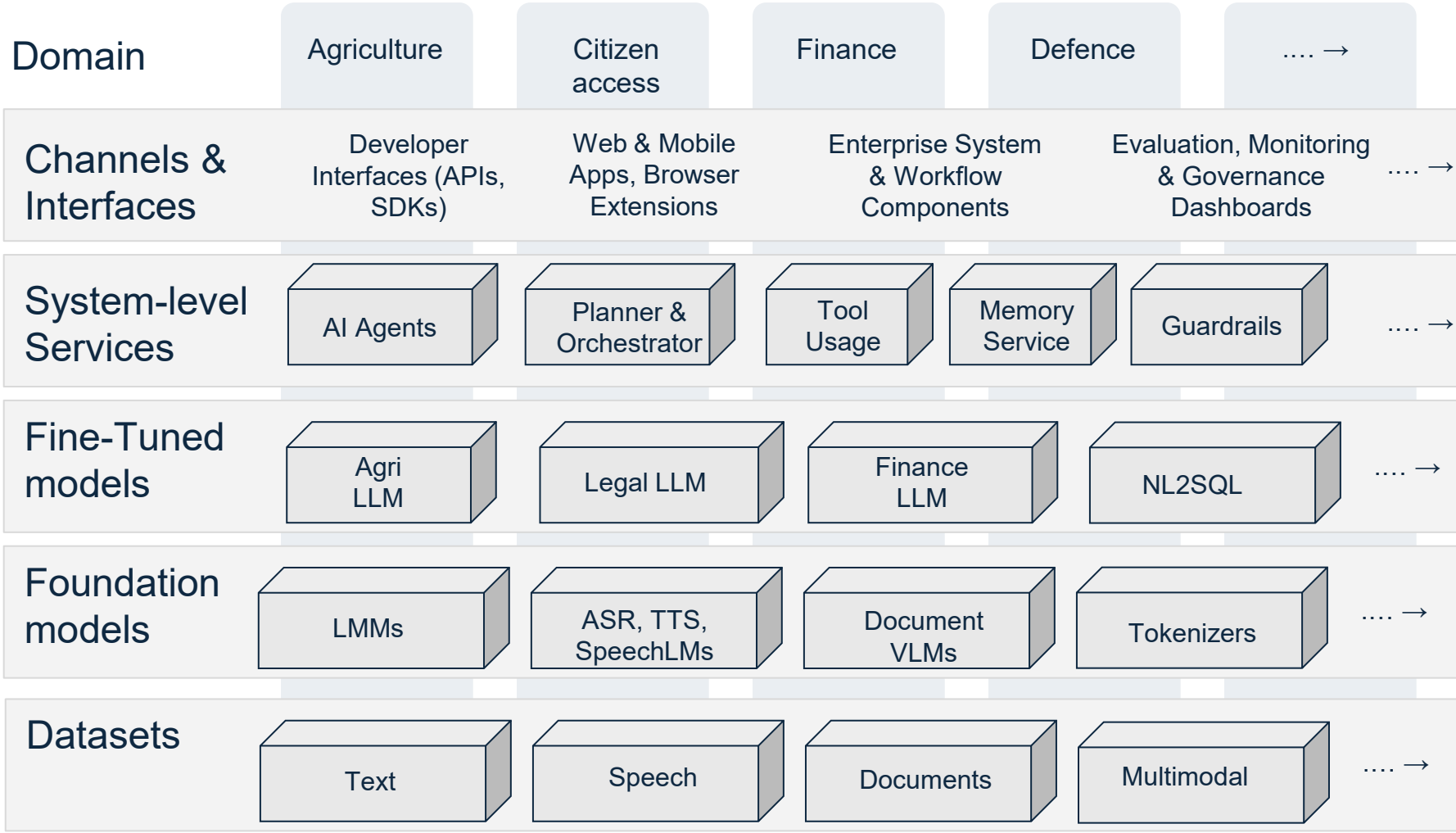
- **For profit companies - Enterprise and Startup**
  - Eg: ICICI Bank, PFL, Godrej, Avalon Consulting, BI CXO, Square Yards, Apollo Ayurvaid, Bosch, Rein Labs, ETV, Tutorials Point,
- **AI and Technology Development partnerships**
  - IBM Research, NVIDIA, …

Academic Outreach and Research

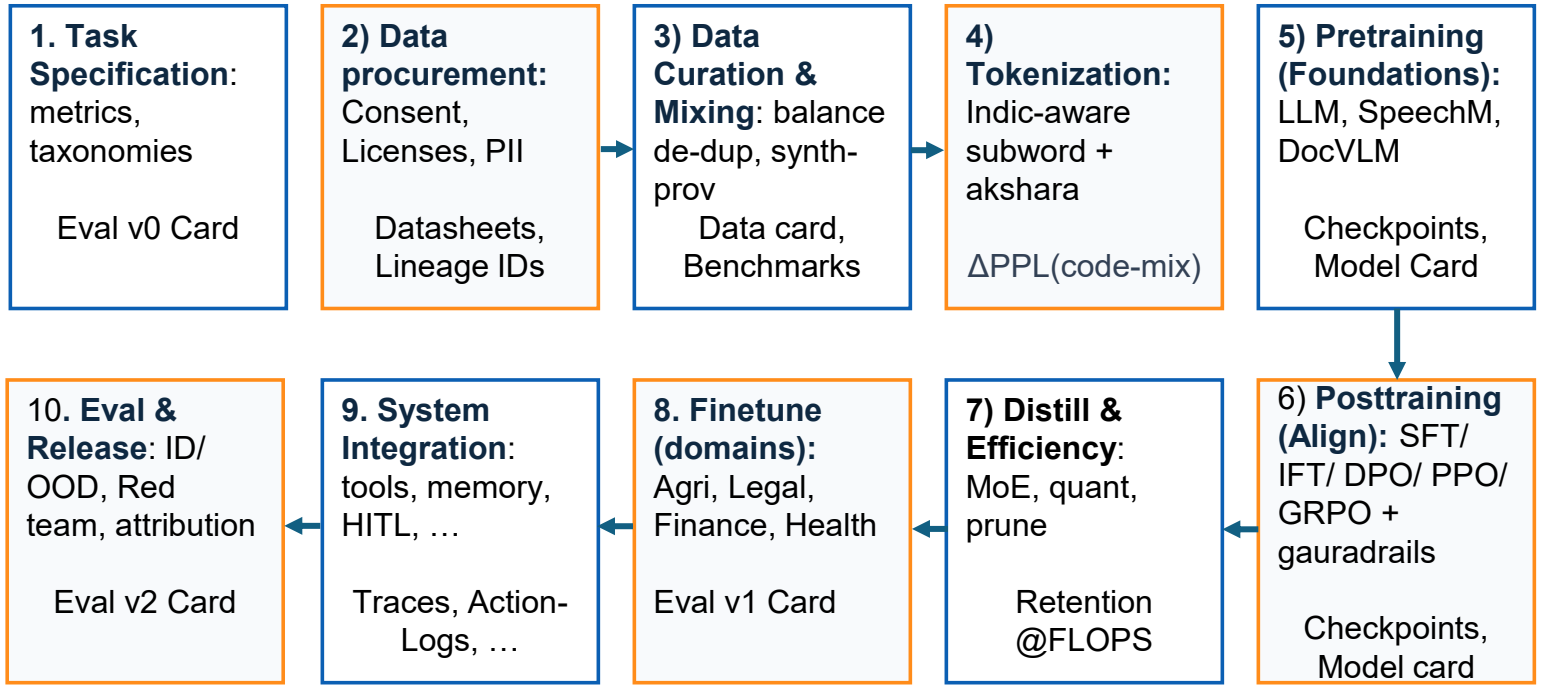
- Partnering with 10+ institutes to provide **50+ PhD, MTech and BTech internships.**
- 7 IITs /IIMs/ IIITs have **compute hardware** purchased for their campus
- **15+ papers** from BharatGen consortium have been accepted at top fora in a short span of 8 months
- Conducted **2 workshops for students** introducing them to AI and innovation.
- Launched **nationwide hackathon** for students.



Scope of Activities


Domain
Agriculture
Citizen access
Finance
Defence
.... →
Channels & Interfaces
Developer Interfaces (APIs, SDKs)
Web & Mobile Apps, Browser Extensions
Enterprise System & Workflow Components
Evaluation, Monitoring & Governance Dashboards
.... →
System-level Services
AI Agents
Planner & Orchestrator
Tool Usage
Memory Service
Guardrails
.... →
Fine-Tuned models
Agri LLM
Legal LLM
Finance LLM
NL2SQL
.... →
Foundation models
LMMs
ASR, TTS, SpeechLMs
Document VLMs
Tokenizers
.... →
Datasets
Text
Speech
Documents
Multimodal
.... →

The Build Pipeline

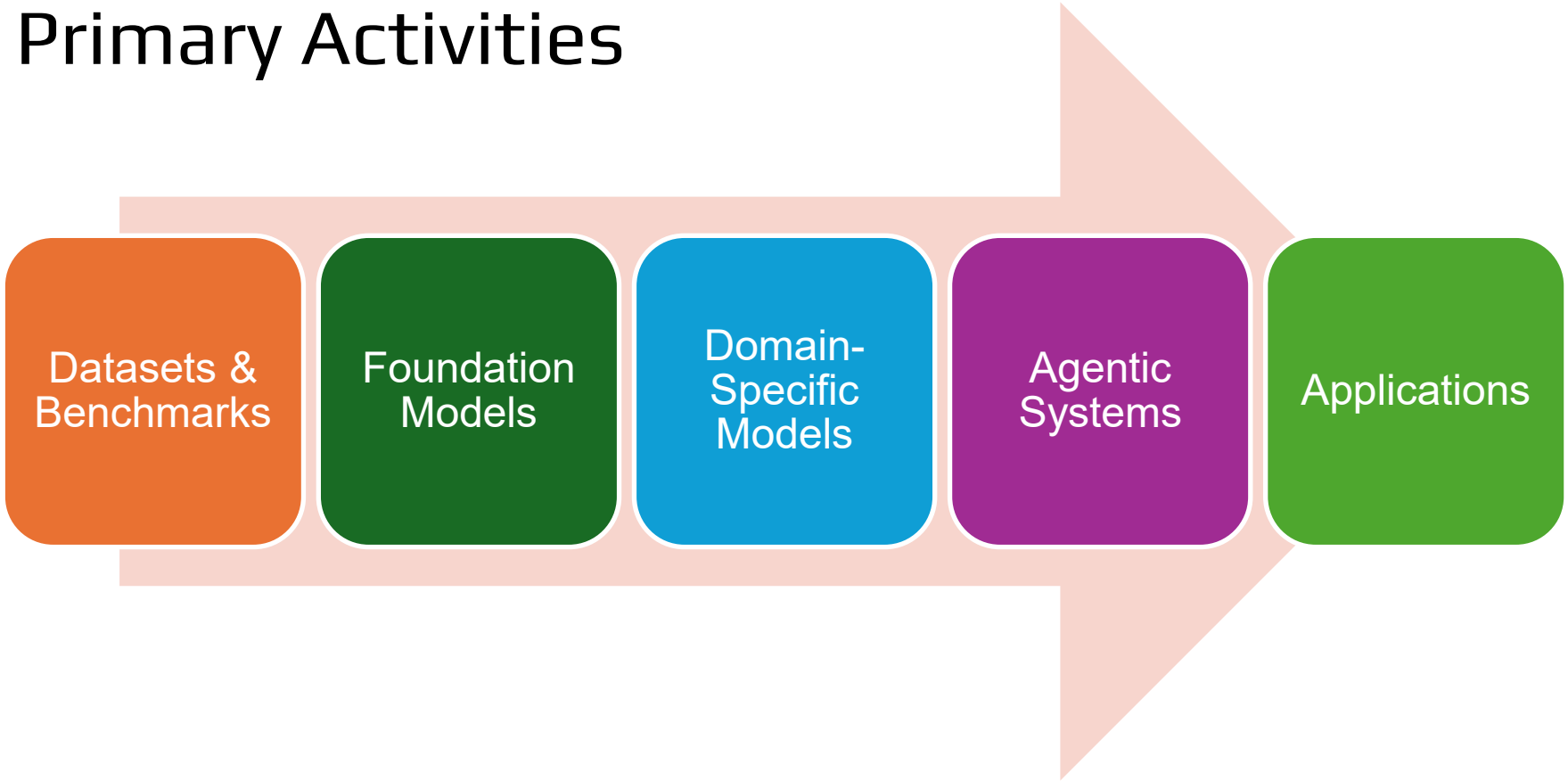


Metrics

- Lineage
- Coverage %
- Robustness Gap
- ΔPPL (code-mix)
- WER / EM-F1 Δ
- Retention @FLOPs
- Joules/ Infer
- p50/ p95 latency

- Lineage & provenance logging end-to-end.

Primary Activities



1st Set of (Small) Foundation Models

A range of **foundational models from 500M to 7 Billion** parameters with 5T tokens

On June 2nd, 2025, BharatGen officially launched

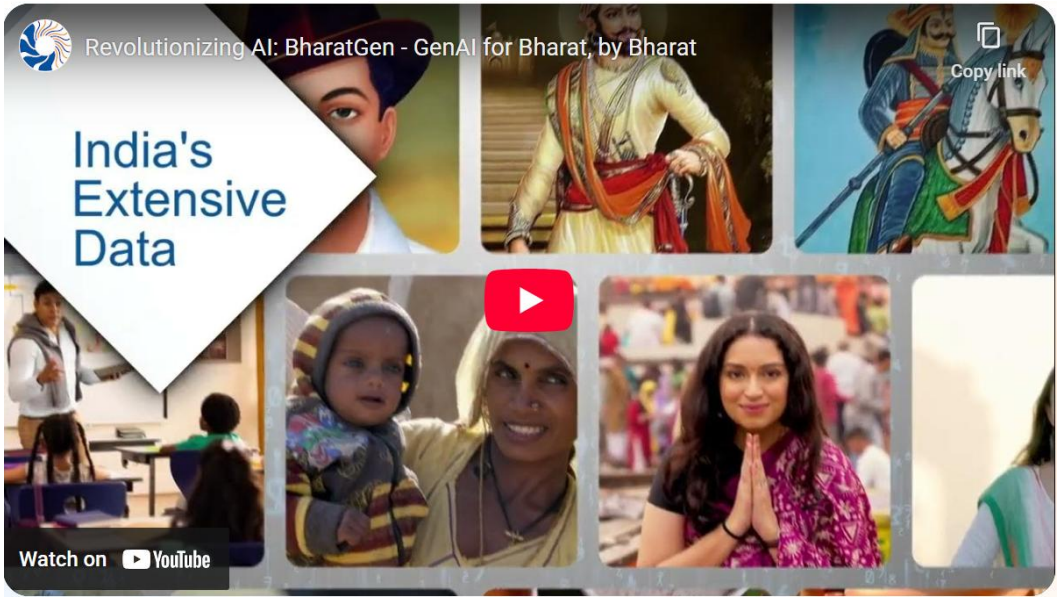
- **Param 1** - 2.9B LLM (pretrained on 7.5T tokens with 33.4% Indian data) and its SFT version (805K samples india centric).
- **TTS models** (**Sooktam**) (150M) for 9 Indic languages,
- **Speaker-conditioned TTS** (**Sooktam**) for 5 languages,
- **ASR (Shrutam)** with 30M parameters.
- **Patram**, India's first vision-language document model (7B), trained on 2.5B tokens.

All models are on Huggingface and on AIKosh:
Huggingface: https://huggingface.co/bharatgenai
AIKosh: https://aikosh.indiaai.gov.in/home/models

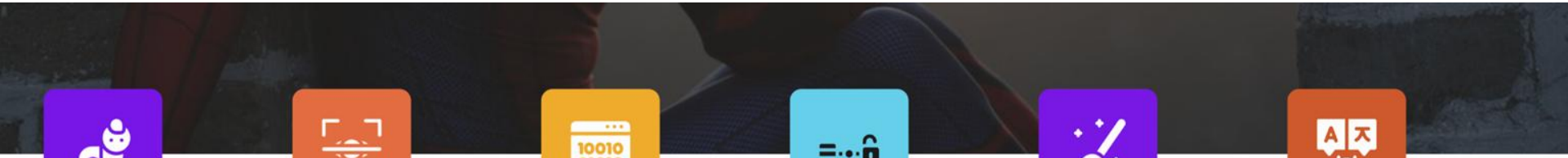
Multilingual Datasets and Benchmarks

Bharat Data Sagar (BDS)

- Initiative for capturing India’s rich and diverse cultural, linguistic, and heritage data.
- Comprehensive datasets across text, speech, images, and other media.
- Representing India’s 22+ official languages and regional dialects.



Multilingual Data Pipeline



Web-Text Pipeline

A robust web-text pipeline collects and cleans millions of documents in multiple Indian languages, ensuring a vast and diverse dataset for language modeling.

Multimodal Extraction

Live multimodal extraction leverages OCR and vision-language ensembles to convert scanned manuscripts and PDFs into accurate, enhanced text suitable for training.

Synthetic Data Generation

Synthetic data generation has started for low-resource languages and coding tasks, with sandbox environments ensuring output correctness and data reliability.

Local Data Partnerships

Initial local partnerships provide offline regional journals and transcripts that are processed into high-quality training data, enriching the corpus with authentic offline sources.

Text Cleaning and Enhancement

Text extracted from diverse sources undergoes rigorous cleaning and enhancement to improve accuracy and usability for model pretraining.

Growing Multilingual Corpus

These integrated components combine to form a continuously expanding, multilingual text corpus, laying the foundation for a strong, inclusive foundation model.

Indic Benchmark Initiative [NAACL 2025]

[20+ benchmarks released on https://aikosha.indiaai.gov.in/ published also at ACL, Interspeech, etc]

Example: IndiaQA Benchmark

Datasets	As	Bn	Gu	Hi	Kn	Ml	Mr	Od	Pa	Ta	Te
Hindi Squad	3099	3107	3371	4734	3068	2926	3165	3079	3469	2743	2955
NQ Open	1462	1483	1570	1842	1447	1420	1511	1451	1570	1331	1420
Chaii	339	351	394	746	351	328	373	305	388	325	361
Indic QA	1789	1763	1369	1547	1517	1589	1604	1680	1542	1804	1734
XSquad	1190	1190	1190	1190	1190	1190	1190	1190	1190	1190	1190
XORQA	537	538	532	537	534	533	529	529	531	537	538
MLQA	2362	2403	2718	4918	2299	2128	2433	2370	2730	2129	2291
Synthetic MCQA*	1741	1662	2162	3802	1618	1248	1807	1753	2326	1150	1416
MS Marco*	29724	30089	31741	35735	29212	28528	30180	30073	32032	27197	28995
Llama Index*	1158	1312	1333	1384	1310	1250	1316	1263	1175	1258	1306

Dataset Creation:

- Created **10 different datasets** in 11 Indic languages.
- Most datasets are **extractive**, while some are **abstractive** (e.g., syn_data, llama_index, mMarco).

Translation Quality Assurance:

- Translation using in-house Udaan translation ecosystem
- To ensure translation quality, we back-translated the translated data to the original language.
- Calculated **CHRF** and **CHRF++** scores for validation.

Abstractive Data Generation:

- For **syn_data**, we sampled paragraphs from Wikipedia across different domains.
- Humanly Verified

Model Evaluation:

- Evaluated the performance of various proprietary and open-source **large language models (LLMs)**.
- Focus on LLMs supporting **multilingual Indic languages**, including both base and instruction-fine-tuned variants.
- Also Reported Zero and few shot numbers.

Synthetic Data Generation Framework

ARISE: Iterative Rule Induction and Synthetic Data Generation [NAACL 25]

Synthetic Data Generation: Virtual Persona & Data Programming

Quality-Filtered Dataset Transformation:

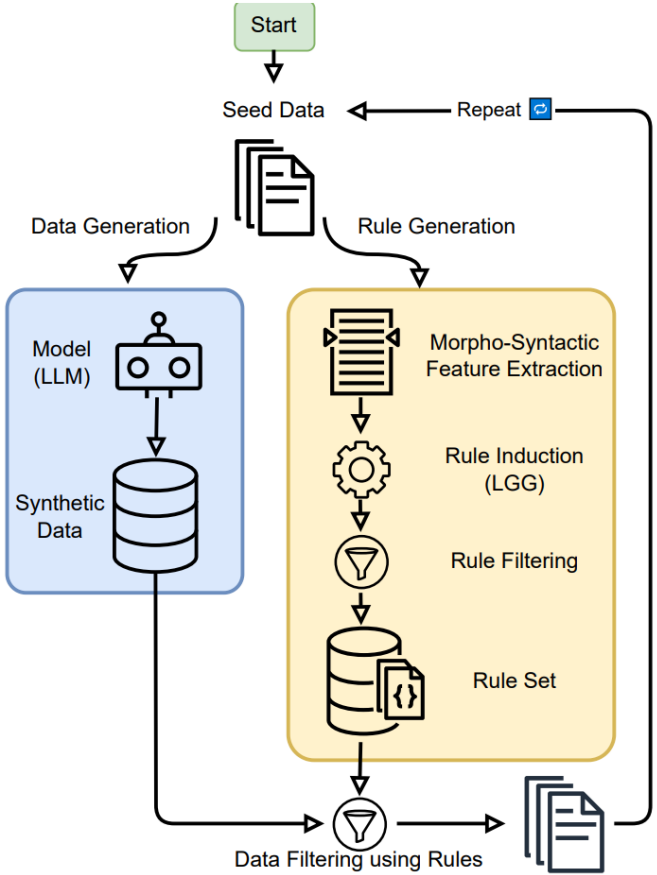
- Converts 250B+ English tokens into Indianized data
- Applies quality filters, Indianisation, and multilingual translation
- Robust QA pipeline ensures relevance and accuracy

Factual Data Generation:

- Use personas to answer factual questions (e.g., What/How/Why...)
- Output is synthetic but precise, grounded in factual templates

Curation:

- Remove toxic content and deduplication due to synthetic control
- Filters for repeating n-grams, symbol-heavy responses, and hallucinat
- Decontamination to ensure benchmark dataset integrity



Synthetic Data Generation Framework and Overview

Objective:

To generate high-quality synthetic data that will be used to create culturally diverse, linguistically rich, and contextually relevant data for improving LLM training.

Why Synthetic Data Matters:

- Scales LLMs efficiently with diverse, bias-controlled datasets
- Addresses data scarcity, privacy, and annotation costs
- Improves generalization, robustness, and real-world performance

Core Strategies:

- Virtual Persona Engagement
- Quality-Filtered Dataset Transformation
- Factual Data Generation

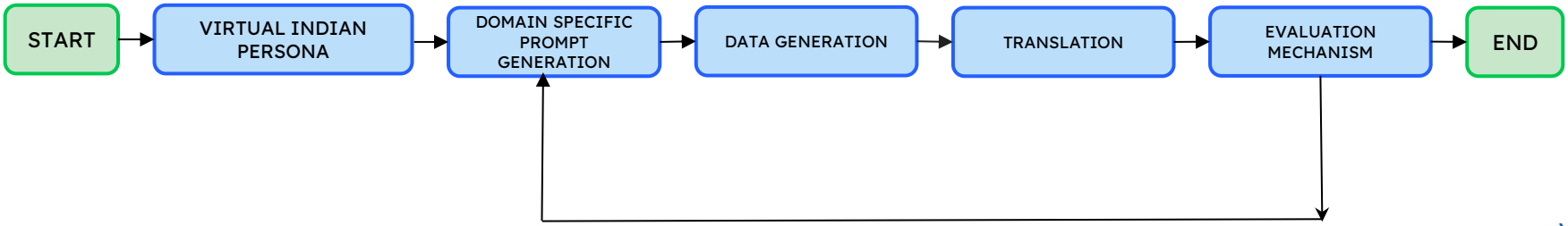
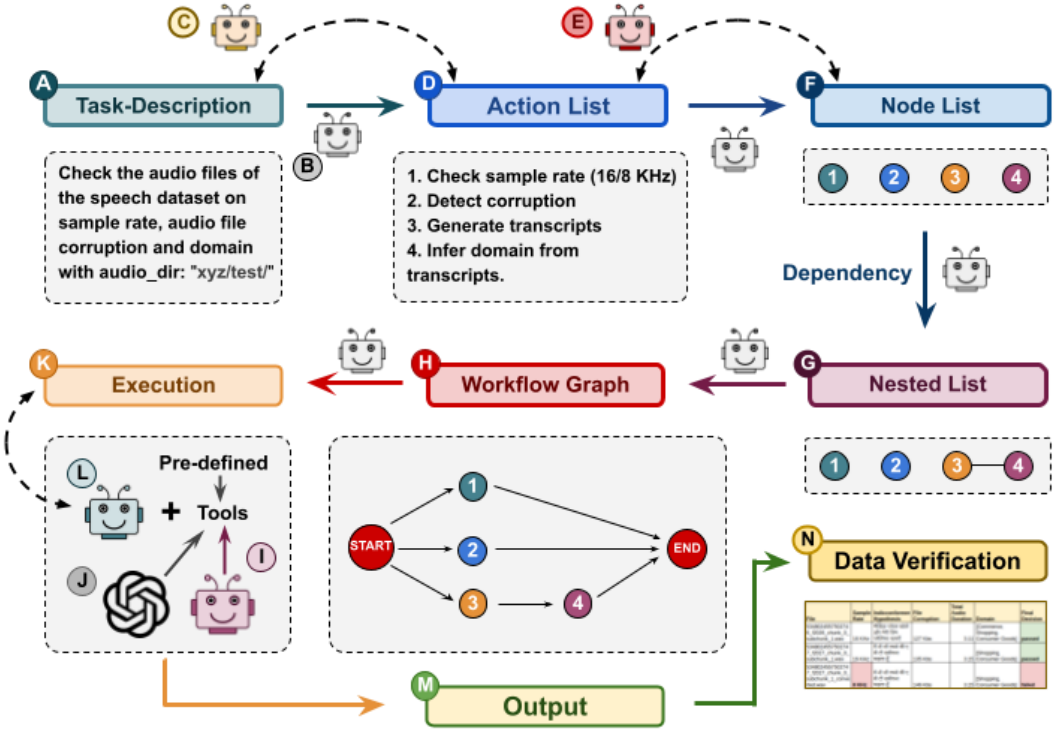


Fig: Broad strategy for persona-based Synthetic data generation

SpeechQC-Agent: A Natural Language Driven Multi-Agent System for Speech Dataset Quality (Paper in preparation)



- Generate **speech dataset verification** workflows.
- Accepts **natural language** task descriptions.
- Converts them into structured **task graphs.**
- LLM-coordinated **multi-agent system.**
- Executes them via **audio**, **transcript**, and **metadata** validation agents

[Video: https://youtu.be/iWoBWmtk338]

(Small) Foundation Models

1st Set of (Small) Foundation Models

A range of **foundational models from 500M to 7 Billion** parameters with 5T tokens

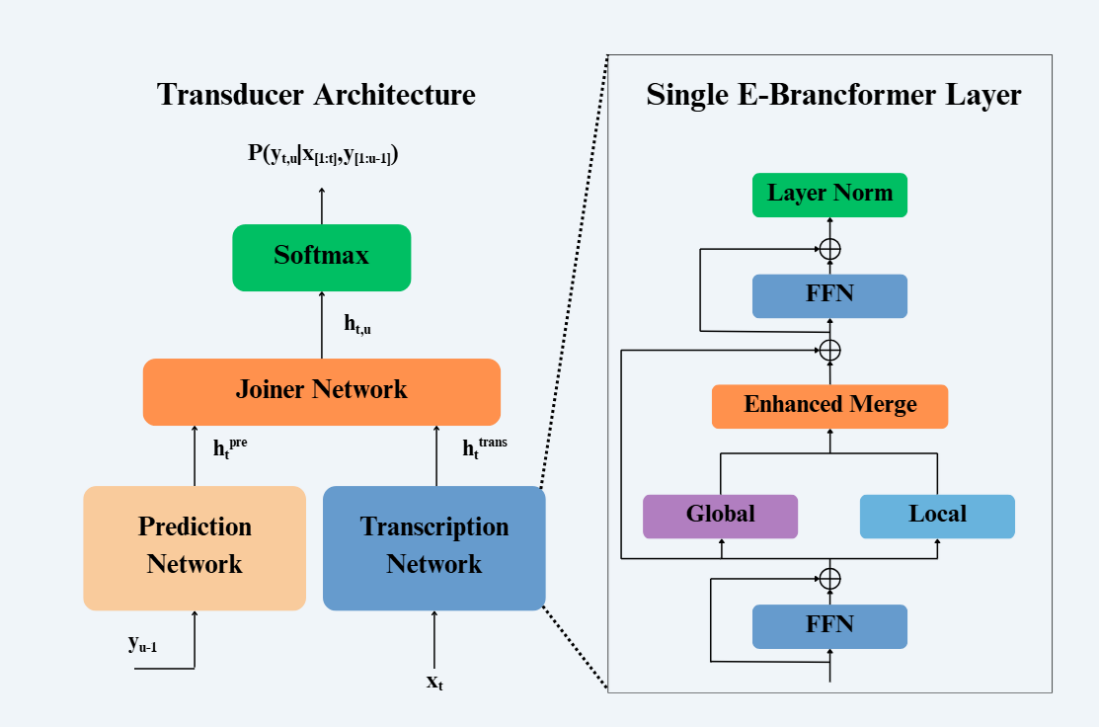
On June 2nd, 2025, BharatGen officially launched

- **Param 1** - 2.9B LLM (pretrained on 7.5T tokens with 33.4% Indian data) and its SFT version (805K samples india centric).
- **TTS models** (**Sooktam**) (150M) for 9 Indic languages,
- **Speaker-conditioned TTS** (**Sooktam**) for 5 languages,
- **ASR (Shrutam)** with 30M parameters.
- **Patram**, India's first vision-language document model (7B), trained on 2.5B tokens.

All models are on Huggingface and on AIKosh:
Huggingface: https://huggingface.co/bharatgenai
AIKosh: https://aikosh.indiaai.gov.in/home/models

(Small) Foundation Models - Speech

ASR Architecture & Coverage of Indic Languages



Language	Durations (in hrs)
Hindi	1655:08:07
Telugu	1472:50:37
Bengali	1363:36:08
Tamil	1317:23:18
Marathi	1307:29:06
Malayalam	670:31:55
Kannada	624:18:31
Punjabi	566:58:02
Assamese	453:06:48
Gujarati	413:23:56
Kashmiri	393:18:59
Urdu	392:53:59
Dogri	362:54:57
Bodo	302:29:24
Manipuri	197:23:28
Nepali	185:03:33
Maithili	93:12:18
Oriya	68:07:25

BharatGen's ASR Model Architecture:

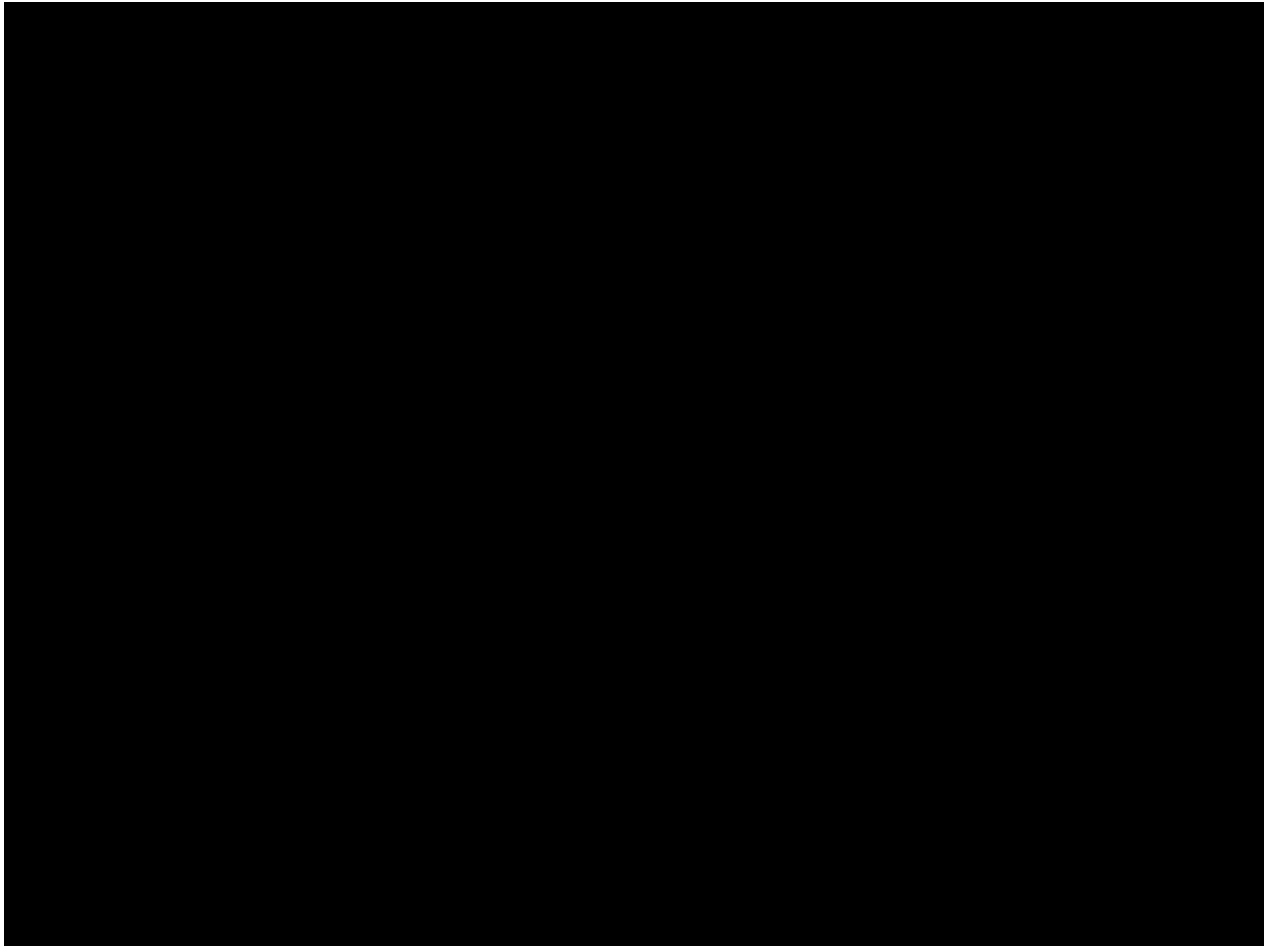
- Hybrid CTC-Transducer model with E-branchformer Encoder
- LM rescoring with LM weight of 0.2 to enhance ASR performance

ASR Demo - https://skilled-charmed-elk.ngrok-free.app/

ASR Benchmark Performance

Model	Parameters	Common Voice	Fleurs	Gram Vaani	IndicTTS	Kathbath	Kathbath Noisy	Mucs	Average Vistar
BharatGen	140M	10.24	12.82	23.57	6.01	8.87	9.97	9.59	**11.58**
IndicConformer(RNNT)	120M	11.32	12.20	24.18	7.48	8.84	9.87	8.90	11.83
AWS		12.85	11.47	21.67	13.94	8.85	9.18	8.70	12.38
IndicConformer(CTC)	120M	12.55	12.52	25.08	8.72	9.61	10.88	10.20	12.79
Azure STT		11.64	12.01	27.71	12.86	9.90	10.51	9.81	13.49
11Labs		11.29	10.47	37.18	12.32	7.64	9.77	10.07	14.11
Sarvam		13.77	11.83	27.15	19.90	10.09	11.92	11.26	15.13
IndicWhisper	769M	15.00	11.40	26.80	7.60	10.30	12.00	12.00	13.59
Google CP- STT		20.83	18.25	59.81	18.18	14.21	16.40	17.77	23.64
Nvidia-large	120M	21.20	15.70	42.60	12.20	12.70	14.20	11.80	18.63
Nvidia-medium	23.5M	20.40	19.40	41.30	12.30	14.00	15.60	12.40	19.34
Azure STT	275M	14.60	24.30	42.30	15.20	13.60	15.10	15.10	20.03
IndicWav2vec	317M	20.20	18.30	42.10	15.00	12.20	16.20	22.90	20.99
Google STT	120M	20.80	19.40	59.90	18.30	14.30	16.70	17.80	23.89

ASR Demo - https://skilled-charmed-elk.ngrok-free.app/



ASR Demo - BharatGen Tamil ASR

Roadmap: Speech models

Text to Speech TTS [[36]]

Dec 2025
TTS
11 languages including Hindi, Bengali, Telugu, Marathi, Tamil, Gujarati, Kannada, Malayalam, Punjabi, Urdu, Odia.

Apr 2026
TTS
16 languages. The choice of the additional 5 languages will be based on feasibility of data collection.

Jun 2026
TTS
All 22 scheduled languages will be covered.

Speech to Text STT [[35]]

Sep 2025
STT
4 languages: Hindi, Marathi, Tamil, Bengali — covering all accents and variants.

Dec 2025
STT
Expand to 11: Hindi, Bengali, Telugu, Marathi, Tamil, Gujarati, Kannada, Malayalam, Punjabi, Urdu, Odia.

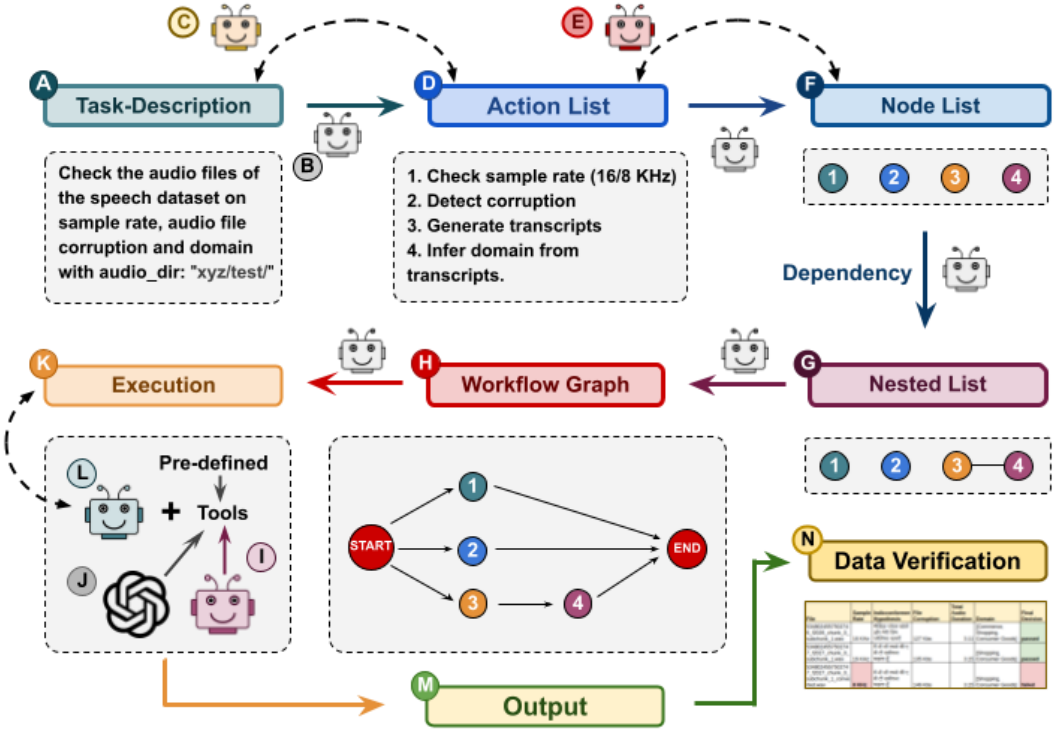
Apr 2026
STT
16 languages. The choice of the additional 5 languages will be based on feasibility of data collection.

Jun 2026
STT
All 22 scheduled languages will be covered.

Speech models: R&D Efforts

- Sourcing speech data
- High quality benchmark creation with Agentic Quality Control: for selecting high quality data across vendors covering diverse accents, conversational speech, with high transcription quality.
- Multilingual ASR models: 4 languages; up to 11 languages by December
- Multilingual Speech Models:

SpeechQC-Agent: A Natural Language Driven Multi-Agent System for Speech Dataset Quality (Paper in preparation)

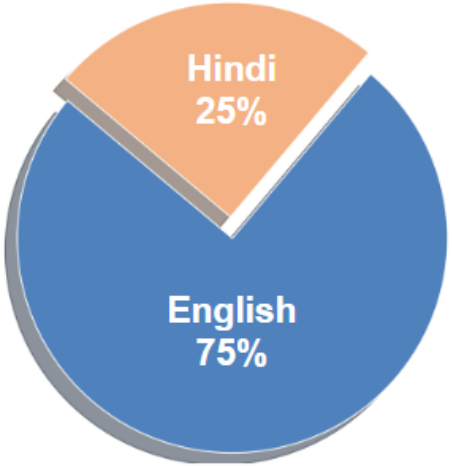


- Generate **speech dataset verification** workflows.
- Accepts **natural language** task descriptions.
- Converts them into structured **task graphs.**
- LLM-coordinated **multi-agent system.**
- Executes them via **audio**, **transcript**, and **metadata** validation agents

[Video: https://youtu.be/iWoBWmtk338]

(Small) Foundation Models - Text

BharatGen : Param 1 : 2.9B (PT) model



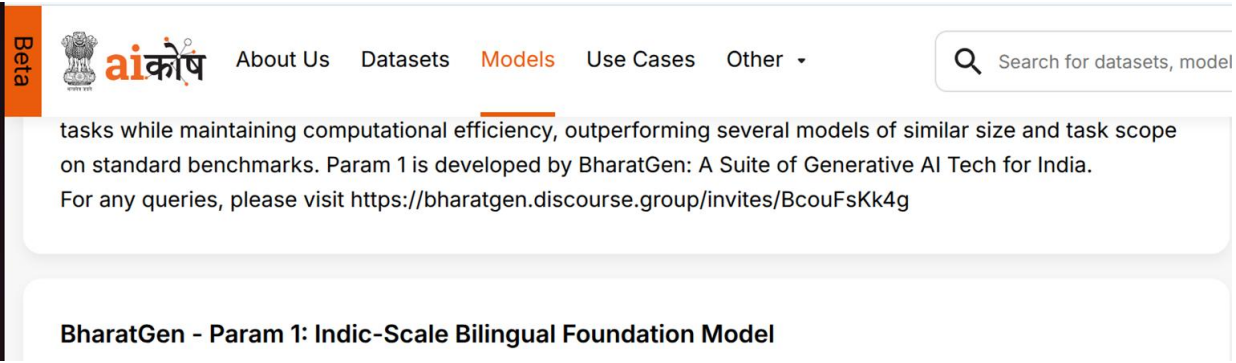
Training Details:

Dataset: **5 Trillion tokens**

- **English**: **3.78T** tokens
- **Hindi**: **1.26T** tokens
- Data Quality: Highly curated with standard filtering and multiple processing steps.
- **Training Setup**: Running on **128 nodes for 2 epochs**
- **Training Duration: 9 days**

Architecture Details:

vocab_size: `256k`
hidden_size: `2048`
intermediate_size: `7168`
max_position_embeddings: `2048`
num_of_attention_heads: `16`
rope_theta: `10000`
num_of_hidden_layers: `32`
num_of_key_value_heads: `8`
attention_function: `SiLU`
attention: `Grouped-query attention`
precision: `bf16-mixed`



https://aikosh.indiaai.gov.in/home/models/details/bharatgen_param_1_indic_scale_bilingual_foundation_model.html

LLM Performance

	Hellaswag (Hi)		MMLU (Hi)		MILU (Hi)	MILU (En)	SANSKRITI
	Zero	Few	Zero	Few			
PARAM-1 2.9B	44.1	45.7	30.7	36.1	30.17	36.3	60.15
QWEN-3B	32.9	32.80	38.32	40.40	33.6	49.84	69.72
SARVAM-2B	42.9	43.8	42.4	41.4	28.48	32.12	52.61
GEMMA-2B	38.6	39.1	30.0	35.8	29.17	44.65	69.76
LLAMA-3B	40.0	40.6	35.0	37.5	29.36	37.63	55.47
GRANITE-2B	31.0	31.1	29.0	30.61	26.06	36.08	60.95

Table 4 Performance comparison on Indic benchmarks.

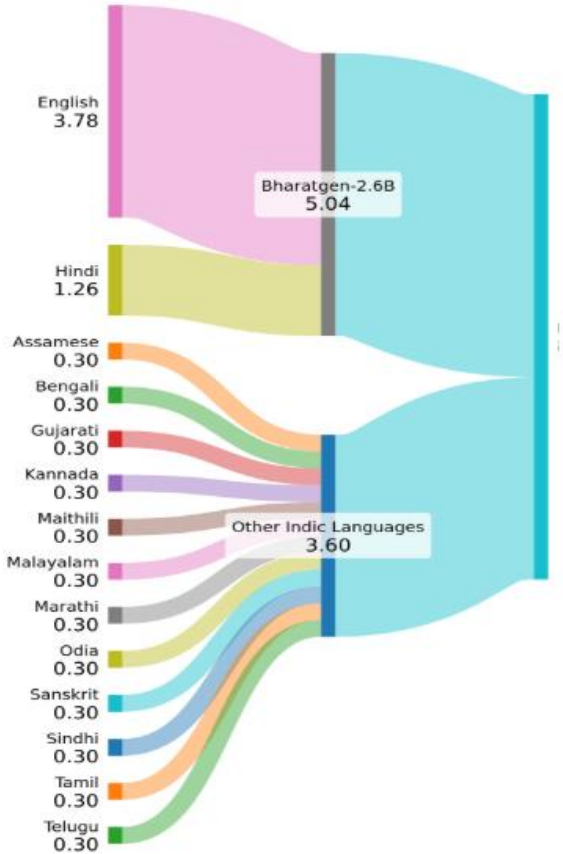
📘 **Sample models. benchmarking and publications**

- aikosh.indiaai.gov.in/home/models/details/bharatgen_param_1_indic_scale_bilingual_foundation_model.html
- huggingface.co/bharatgenai/Param-1-2.9B-Instruct
- **bharatgen.com/param-revolutionizing-ai-for-india/**
- PARAM-1 BharatGen 2.9B Model
- huggingface.co/datasets/bharatgenai/BhashaBench-Krishi

📘 **Sample models. benchmarking and publications**

- The Art of Breaking Words: Rethinking Multilingual Tokenizer Design
- Multilingual Tokenization through the Lens of Indian Languages: Challenges and Insights
- MorphTok: Morphologically Grounded Tokenization for Indian Languages
- huggingface.co/bharatgenai/AgriParam etc….

Ongoing: 7B Mixture of Experts (MoE) model



ing Dataset: 8.6 trillion token

Indian Languages: 3.6 trillion tokens across **16 Indian languages**:

- Assamese (asm_Beng)
- Bengali (ben_Beng)
- Gujarati (guj_Gujr)
- Hindi (hin_Deva)
- Kannada (kan_Knda)
- Maithili (mai_Deva)
- Malayalam (mal_Mlym)
- Marathi (mar_Deva)
- Odia (ory_Orya)
- Sanskrit (san_Deva)
- Sindhi (snd_Deva) (Devanagari)
- Tamil (tam_Taml)
- Telugu (tel_Telu)

English-Hindi Dataset: 5 trillion tokens

- **English**: 75% (3.78 trillion tokens)
- **Hindi**: 25% (1.26 trillion tokens)

Data Quality: Highly curated with standard filtering and multiple processing steps.

Architecture Details:

vocab_size : `256k`
model_type : `gpt_dolomite`
hidden_size : `1536`
num_layers : `40`
max_position_embeddings : `4096`
rope_theta : `10000`
sequence_mixer_blocks :
    type : `softmax_attention`
    num_attention_heads : `12`
    num_key_value_heads : `4`
mlp_blocks:
    type: `MoE`
    num_experts: `64`
    num_experts_per_tok: `8`
    intermediate_size: `512`
    activation_function: `swiglu`

(Small) Foundation Models – Document VLMs

Addressing the Indian Document Challenge

22+ Languages

Vast linguistic diversity creating barriers to unified data processing.

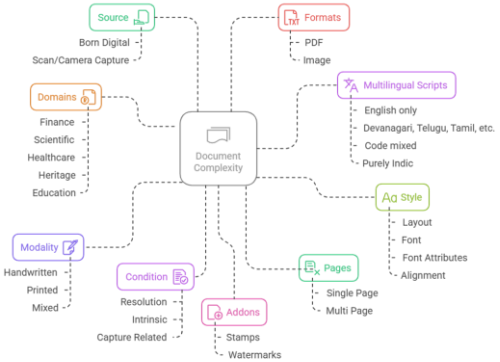
Complex Layouts

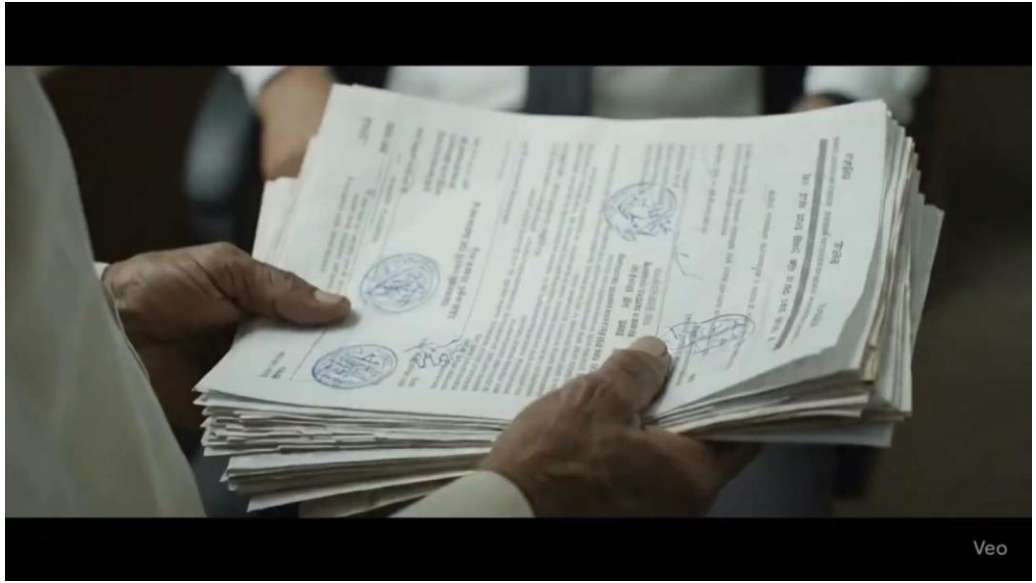
Multi-page, non-standard formats in official and legal documents.

Usecase Sectors

Governance	Finance	Healthcare	Legal
Defense	Agriculture	Education	Public Service

*Foreign VLMs **fail** on Indian scripts, regional variations, and official document formats.*





Patram-India-centric Vision Foundation Model from Scratch

Multi-page Layout Understanding

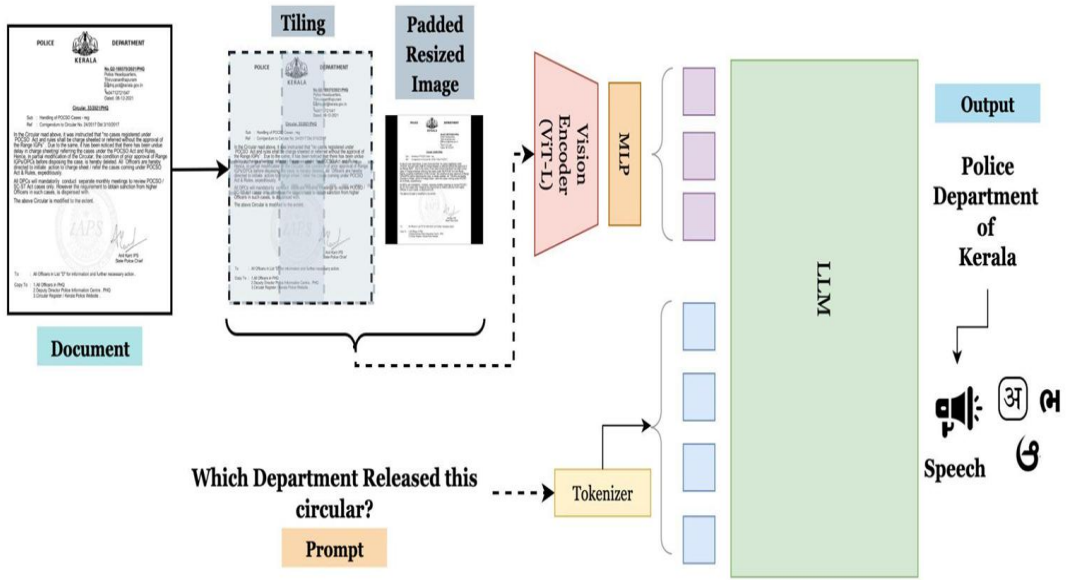
Comprehends complex structures across multiple pages, essential for reports and contracts.

End-to-End Models

Integrated pipeline for OCR, VQA, Classification, and Summarization.

Multilingual & Domain-Specific

Fine-tuned for Finance, Legal, Healthcare, and Governance in 22 Indian languages.



VLM Performance

Model	Overall	DocVQA	Patram-Bench	VisualMRC
Qwen2.5-VL-7B-Instruct	0.8759	0.8722	0.6816	**0.9169**
Patram-7B-Instruct	0.8692	**0.9373**	0.6389	0.8598
Gemma-3-12B-IT	0.8556	0.8451	0.6349	0.9069
InternVL3-9B	0.7865	0.8681	0.6888	0.7405
DeepSeek-VL2	0.7581	0.8739	0.5089	0.7144
Molmo-7B-D-0924	0.6820	0.6314	0.5018	0.7577
DocOwl2	0.6776	0.7391	0.4569	0.5868
Molmo-7B-O-0924	0.6654	0.6314	0.5018	0.7577
Llava-v1.6-mistral-7b	0.2577	0.1263	0.4134	0.3889

Sample models. benchmarking & publications

- huggingface.co/bharatgenai/patram-7b-instruct
- DRISHTIKON: Visual Grounding at Multiple Granularities in Documents
- DORM: DocVQA via OCR and RAG for Multilinguality
- Tables Decoded: DELTA for Structure, TARQA for Understanding
- And so on..

Domain-Specific Models & Benchmarks

Building Domain-Specific AI for India

Comprehensive Benchmarks & Models for 4 Important Indian centric domains Agriculture, Finance, Ayurveda, Legal

Fine-tune Param1-2.9B on Domain specific India-centric Dataset

AgriParam - Indian agriculture, crops, policies, agri-business
Dataset Highlights - 17k passages → 2M Q&A, 6M dialogues (EN+HI)

FinanceParam - Indian finance, taxation, banking, insurance, investments
Dataset Highlights - 25k+ passages → 9M Q&A, 8M dialogues (EN+HI)

AyurParam - Ayurveda texts, clinical knowledge, research
Dataset Highlights - 1000+ texts (~54.5M words) → 4.8M Q&A/dialogues (EN+HI)

LegalParam - Indian legal texts, case law, regulations, judgments, policies
Dataset Highlights - 30k+ legal documents → 5M Q&A/dialogues (EN+HI)

Benchmark Releases

- BhashaBench-Krishi 🌾
- BhashaBench-Ayur 🌿
- BhashaBench-Finance 📈
- BhashaBench-Legal ⚖️

Model Release

- AgriParam
- AyurParam
- FinanceParam
- LegalParam

Releasing Soon

- Param2 - 7B MoE (Model)

Benchmarks for Agriculture

BhashaBench - Krishi

First large-scale benchmark for Indian agricultural knowledge

English & Hindi

25+ agricultural and allied science domains

Region-specific, actionable knowledge for farmers

MCQ, Assertion-Reasoning, Match Column, Rearrange, Fill in Blanks

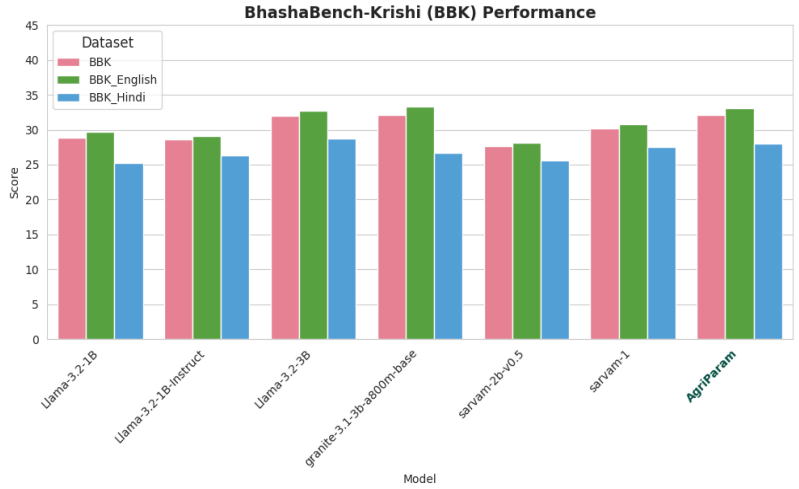
Models Evaluated: 29+ (GPT-4o, Qwen3-235B, open-source LLMs)

Top Accuracy:
English: 70%+
Hindi: 60–65%

Challenges: Hard questions & non-MCQ formats

AgriParam

- ➔ **Trained on India-centric agricultural data (12M+ samples)**
- ➔ **Bilingual (English + Hindi)**, context-aware responses for farmers
- ➔ **Covers crop practices, rural advisory, policy schemes, and agri-research insights.**
- ➔ **Multi-turn conversation support**
- ➔ **Competitive performance** on BhashaBench-Krishi



Financial Benchmarks

BhashaBench - Finance

First comprehensive benchmark designed to evaluate AI models on Indian financial knowledge and practices

English & Hindi

30+ financial and allied domains

Practical, regulation-aware, India-specific financial knowledge.

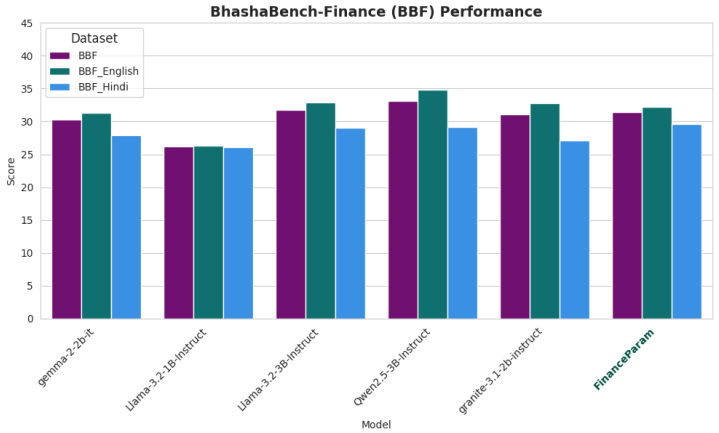
Multiple Choice, Assertion-Reasoning, Match the Column, Rearrange the Sequence, Fill in the Blanks, Reading Comprehension, Essay

Models Evaluated: 40+ (Gemma, Llama, Qwen, & specialized financial AI systems.)

Top Accuracy: 70%+

FinanceParam

- ➔ **Fine-tuned on 25,000+ finance-focused passages from Indian sources.**
- ➔ **Bilingual (English + Hindi), responses grounded in Indian finance, regulatory frameworks, and cultural nuances**
- ➔ **Handles queries on personal finance, taxation, banking, investments, and policy guidance.**
- ➔ **Supports multiple query types – Q&A, reasoning, MCQs, and multi-turn advisory.**



Benchmarks for Legal NLP

BhashaBench - Legal

A comprehensive benchmark designed to rigorously evaluate AI models on Indian legal knowledge.

English & Hindi

20+ legal and allied disciplines

Practical, context-rich, jurisdiction-specific legal knowledge

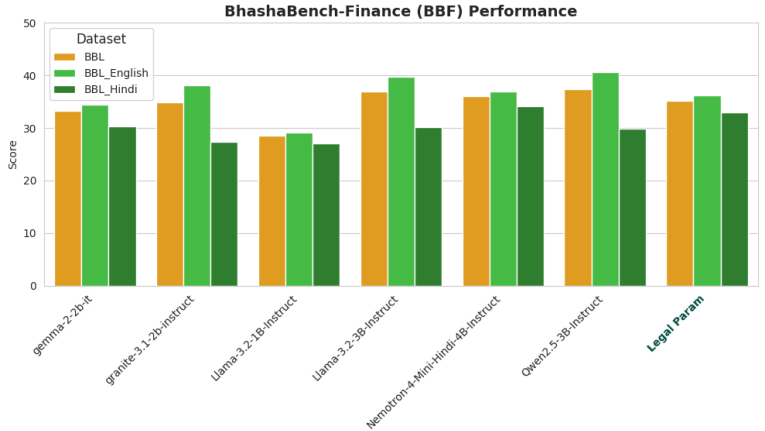
Multiple Choice, Assertion-Reasoning, Match the Column, Rearrange the Sequence, Fill in the Blanks

Models Evaluated: 25+ (GPT-4o, Qwen, DeepSeek & specialized legal models)

Top Accuracy (Larger models): English - 70% to 75% + Hindi - 65-70%

LegalParam

- Domain-specialized large language model build on top of Param-1 on an exhaustive India-centric legal dataset.
- **Bilingual (English + Hindi)**, context-aware responses making legal knowledge more accessible, contextual, and actionable.
- **Delivers accurate, context-aware answers to legal queries while also supporting tasks such as summarizing lengthy legal documents and simplifying complex policy texts**
- **Supports multiple query types** – aiding practitioners with quick references, assisting researchers in exploring legal frameworks, or helping citizens better understand their rights and obligations



Benchmarks for Ayurveda (Traditional Indian Medicine)

BhashaBench - Ayur

First comprehensive benchmark evaluate models on traditional Ayurvedic knowledge and practice

English & Hindi

15+ specialized Ayurvedic disciplines

Traditional knowledge for clinical use, herbal pharmacology, and holistic healthcare

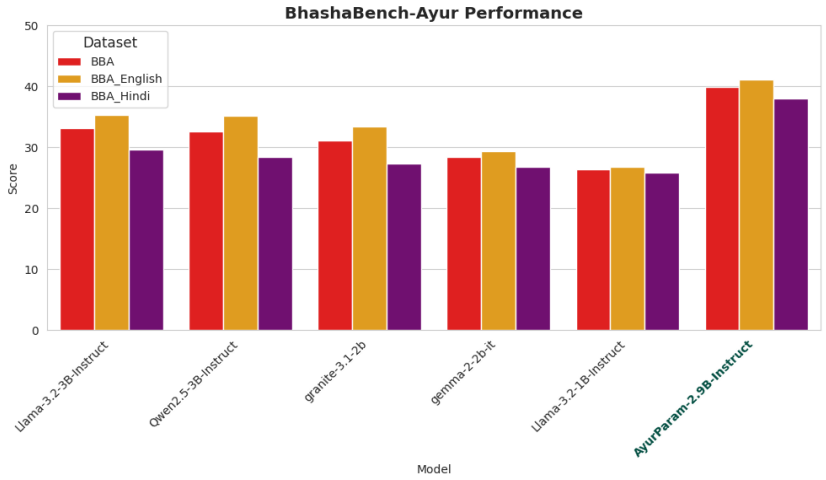
Multiple Choice Questions (MCQ), Fill in the Blanks, Match the Column, Assertion-Reasoning

Models Evaluated: 29 (Qwen3-235B, Gemma-2-27B, Llama-3.1-8B)

Top Accuracy:
English: 60.25%
Hindi: 54.78%

AyurParam

- **Fine-tuned on 1000+ Ayurvedic texts (~54M words).**
- **Bilingual (English + Hindi)**, context-aware responses grounded in Ayurveda.
- **Handles Ayurvedic queries classical text interpretation, clinical guidance, and wellness knowledge.**
- **Supports multiple query types** – Q&A, reasoning, MCQ, and clinical use cases.



Why BharatGen? Sovereignty & Agency

What is Sovereignty?

Sovereignty is the supreme, final, and self-sustaining authority within a political community—the capacity to legislate, adjudicate, and enforce without subordination to any external power—an idea classically defined by Jean Bodin (1576) and elaborated by Thomas Hobbes (1651) as the foundation of legitimate governance.

saor “free noble” (Old Irish sáer) + -ánach = *saoránach* ≈ “one who is free” → *saoránaigh* “free people/citizens.”

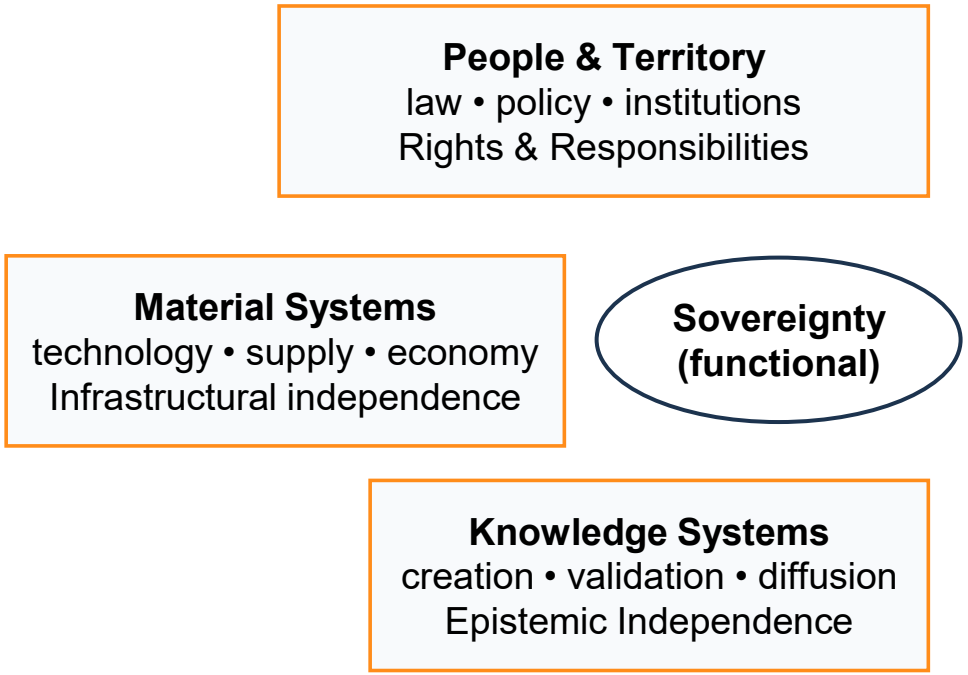
Leviathan, Thomas Hobbes, Ed. Richard Tuck, Cambridge University Press, 1991, pp. 120–126.
Jean Bodin, Six Books of the Commonwealth, trans. M.J. Tooley, Oxford: Basil Blackwell, 1955, p. 25.

Functional Sovereignty

A working definition:

Sovereignty, in its functional sense, denotes the autonomous authority of a nation to govern not only its territory and people but also of material systems -- technological and economic, and knowledge itself -- the creation, validation, and diffusion of understanding that *that underpin the self-governance of its collective life.*

This definition provides a conceptual bridge to *AI sovereignty*.

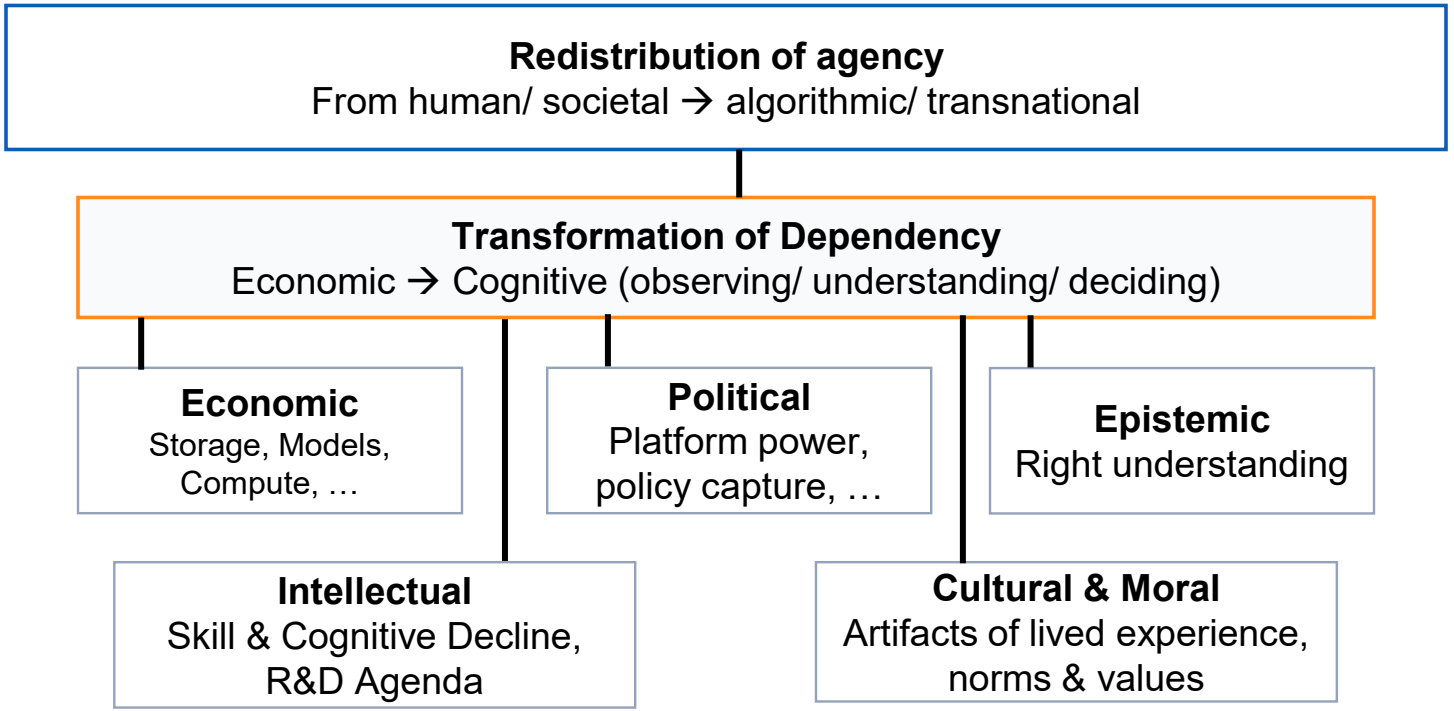


How does AI threaten Sovereignty?

*AI threatens sovereignty not merely by concentrating power externally (economic and thereby, also political), but <u>also</u> by redistributing the **loci of agency** — from the human and the societal to the algorithmic and transnational, it transforms dependency from an economic condition to a **cognitive** one, eroding a people's capacity to know, decide, and imagine for themselves.*

The threat is multifaceted: economic, political, epistemic, cultural, institutional, moral, and intellectual.

How does AI threaten Sovereignty?



What is AI Sovereignty?

A working definition:

AI Sovereignty *is the autonomous capacity of a people and their institutions to design, govern, and evolve AI systems that uphold political independence, economic self-reliance, moral-civilizational integrity, and epistemic freedom; ensuring that the architectures of intelligence serve as instruments of human and societal agency rather than substitutes for them, enabling humans to remain the authors of meaning, judgment, and action in an increasingly algorithmic world.*

AI Sovereignty
People-first, adheres to national controls, aligned with shared scientific norms

Data
Ownership of primary data (observations) of the world

Epistemic
Sources for creating knowledge & understanding

Technical Stack
Data stores, models, compute

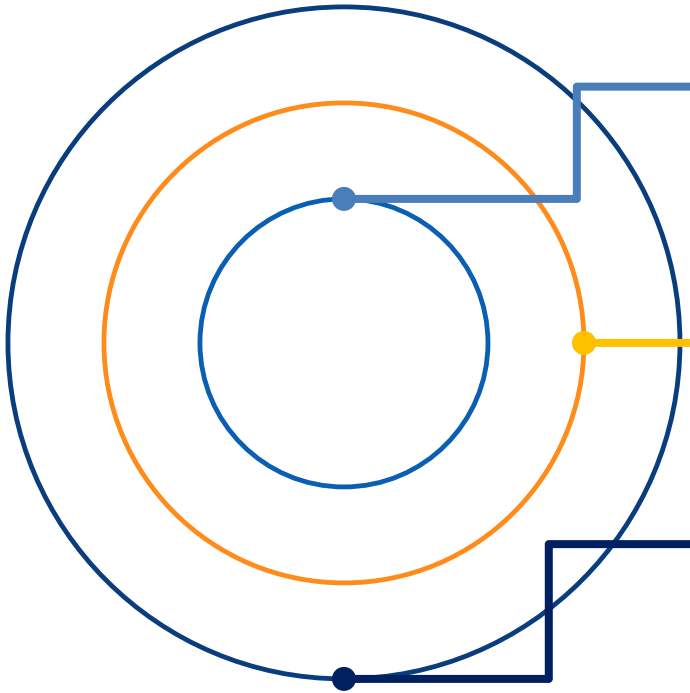
Governance
Cultural norms, societal & human values, scientific norms

Sovereignty and Agency

Political sovereignty ensures **agency** over governance.
Economic sovereignty ensures **agency** over production and livelihood.
Epistemic sovereignty ensures **agency** over thought and understanding.
Moral sovereignty ensures **agency** over conscience and purpose.

AI Sovereignty = the reclamation of agency at all four levels in the age of intelligent systems.

Agency Exists in Hierarchies



Individual
Rights • Choices • Creativity

Societal / National
Language • Culture • Governance

Shared Scientific Commons (Human/ Global)
Open science • Interoperability • Federated evaluation

Consent-based sharing and context-preserving interoperability ensure agency flows inward and outward.

BharatGen's Goals

AI Sovereignty: Beyond Control, Toward Agency

- To design and govern AI systems that serve as **instruments of human and societal agency -** not substitutes for them.
- To ensure that epistemic and cognitive mechanisms - human and artificial - remains accountable to human meaning, dignity, and purpose.

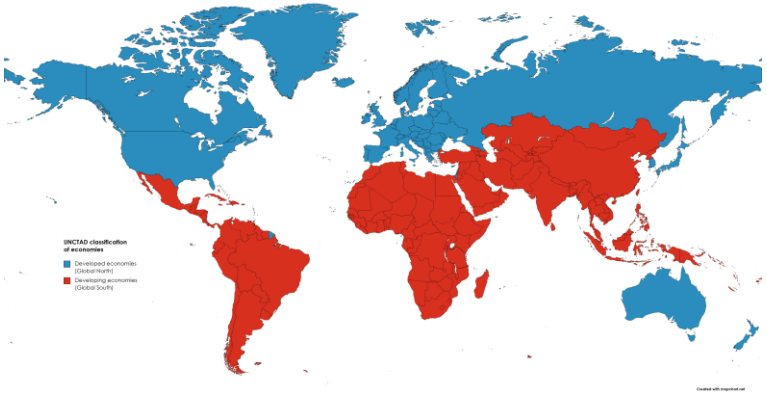
While BharatGen does serve India's sovereignty needs: but recast as above, it creates a framework for all of us to work together towards common goals

Good for one must mean good for all.

Sovereignty for one must enable sovereignty for all.

Call For Action

Call for Action (to the Global South)

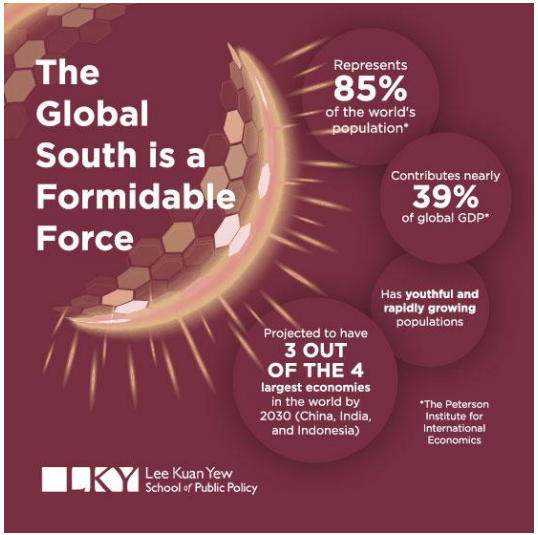


https://en.wikipedia.org/wiki/Global_North_and_Global_South

Why It Matters

- Most humanity lives in regions under-represented in global AI datasets and decision frameworks.
- Without sovereign capability, nations risk epistemic, technological, economic (and political) dependency — **a new digital coloniality**.
- With shared capability, we all can co-author the future of AI and human agency.

Sovereignty for one must enable sovereignty for all.



https://lkyspp.nus.edu.sg/gia/article/what-does-the-rise-of-the-global-south-mean-for-the-world

Call for Action: Frugal-AI Global South Consortium

- **Mission**: reach compute-optimal efficiency without hyperscaler budgets
- **Workstreams**: (a) Open data commons, (b) Synthetic-data toolchains, (c) Equivariance/low-ID modeling, (d) Energy-aware training/inference
- **Governance**: align with global ethics frameworks; regional hubs; public leaderboards with compute budgets per result
- **Invitation**: Let’s collaborate -- create joint data, modeling, evaluation, and low-power inference tracks!

Thank you. Questions!