LARGE LANGUAGE MODELS FOR INDIA

Sundeep Teki, PhD

Leader in AI & Neuroscience

SUNDEEPTEKI.ORG

Slides available at sundeepteki.org/talks

INDUSTRY

- AI Consultant & Coach (sundeepteki.org)
- Head of AI, Docsumo (B2B SaaS AI startup)
- AI Team Lead, Swiggy (B2C unicorn startup)
- AI Scientist, Amazon Alexa AI (Big Tech)

ACADEMIA

- Fellow in Neuroscience, Oxford
- PhD in Neuroscience, UCL
- MSc in Neuroscience, Oxford
- BE in Electronics Engg, DCE

STARTING UP

- Founder, AI Upskilling platform (skills)
- Founder, EduChroma (career mentoring)
- Founder, Timing Research Forum (research)

OUTLINE

1 Large Language Models

2 LLMs for India

3 Business Use Cases

4 Careers & Resources

5 Q & A

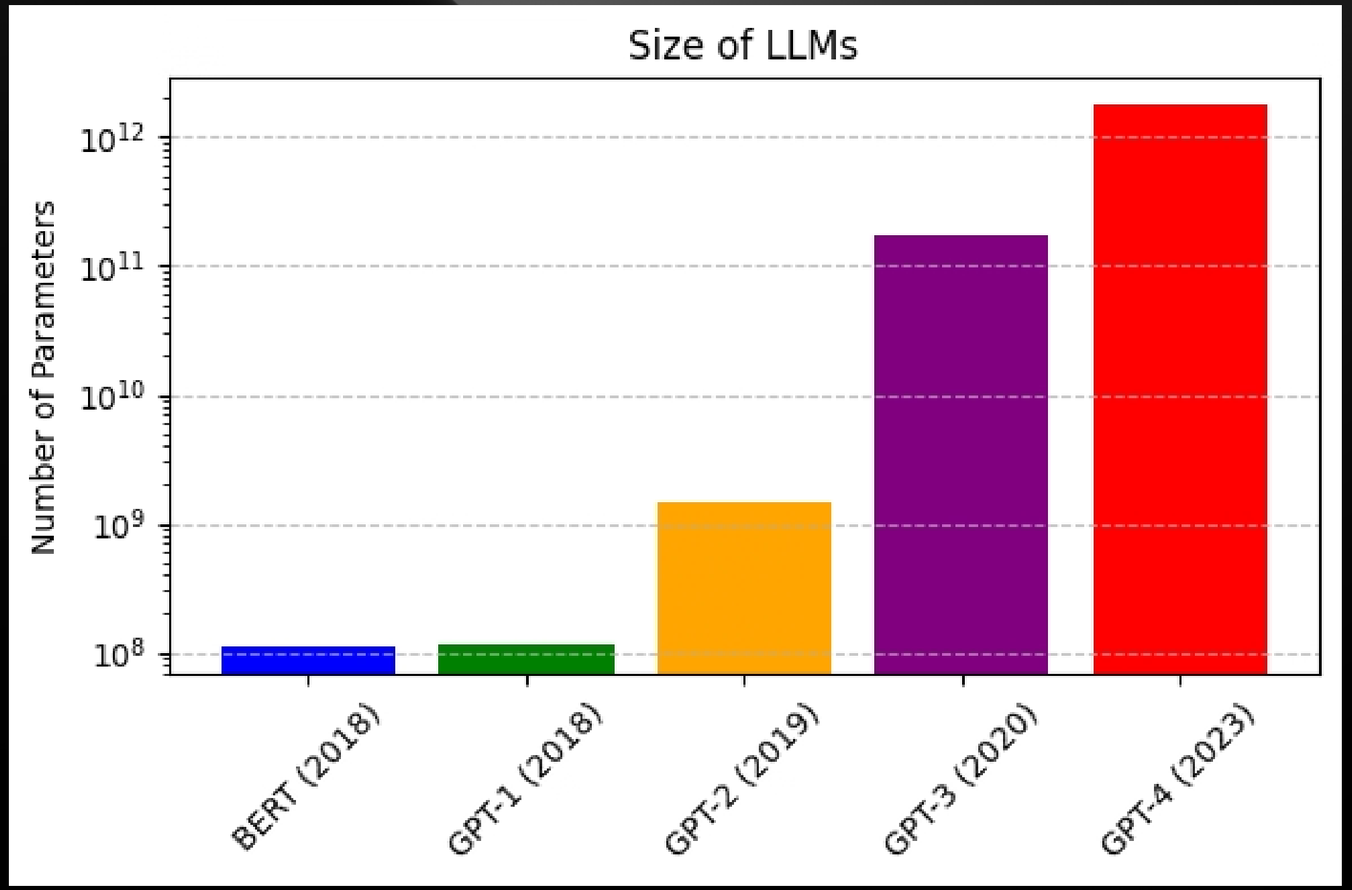
GENERAL-PURPOSE LLMS

- OpenAI's GPT4
- Google's Gemini
- Anthropic's Claude-2
- Meta's LLaMa-2
- Mistral's Mixtral
- Cohere
- Databricks
- Stanford's Alpaca
- TII's Falcon
- LMSyS' Vicuna
- HuggingFace BLOOM
- ...

DOMAIN-SPECIFIC LLMS

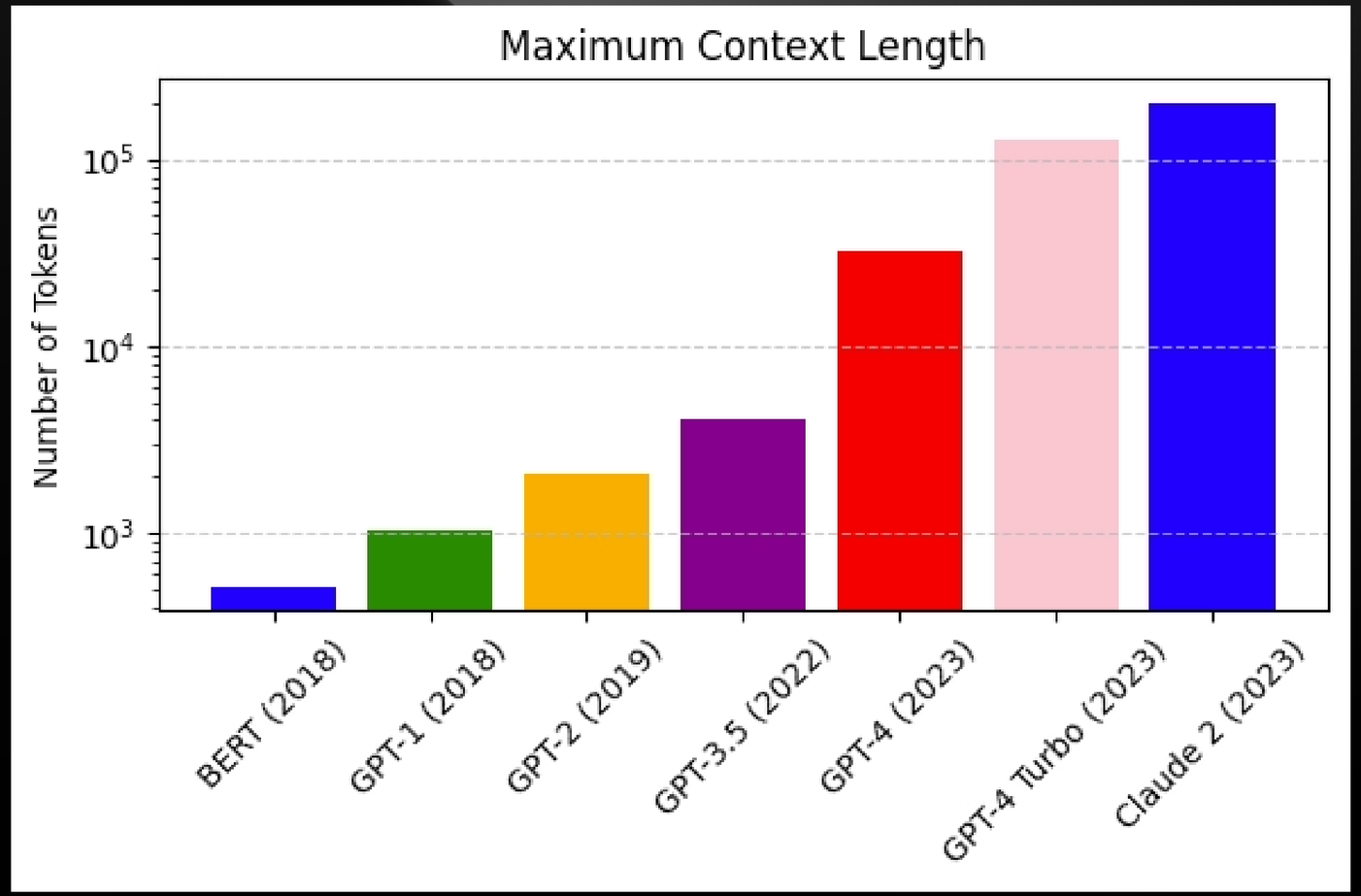
- Bloomberg GPT (Finance)
- Einstein GPT (CRM)
- StarCoder (Code)
- BioGPT (Biomedicine)
- Dhenu (Agriculture)
- ...

EVOLUTION OF LLMS



110 Million (16000x) 1.76 Trillion

EVOLUTION OF LLMS



0.5k (400x) 200k

OUTLINE

1 Large Language Models

2 LLMs for India

3 Business Use Cases

4 Careers & Resources

5 Q & A

WHY BUILD LLMS FOR INDIA?

- Foundational LLMs like LLaMA-2 have <0.2% representation of Indian languages & ~90% English representation where nearly ~10% of India's population can speak English
- The potential of LLMs for India is vast. They can revolutionize sectors like education, healthcare, agriculture, and government services etc.
- Democratize the power of LLMs across India's diverse population, reduce the digital divide and foster greater inclusivity
- To alleviate national data privacy and security concerns for data from critical domains like government, health etc.
- Preserve and promote linguistic diversity and multicultural representation of India and encourage AI solutions for local languages and regions, thereby supporting local economies and use cases
- Promote digitization of non-English data from India to power more digital and AI applications
- Promote building of the entire AI stack and infrastructure in India, supported by local funds and talent

LLMS FOR INDIA

- Bhashini (MEITY, Govt. of India)
- OpenHathi (sarvam.ai)
- Krutrim AI (olakrutrim.com)
- Tamil-LlaMA (independent)
- Telugu LlaMA (independent)
- Malayalam LlaMA (independent)
- Kannada LlaMA (tensoic.com)
- Ambari (cognitivelab.in)
- BharatGPT (corover.ai)
- Project Indus (Tech Mahindra)
- Bharat GPT (Reliance Jio)
- ...

DEMOS

TAMIL-LLAMA:
https://huggingface.co/spaces/abhinand/tamil-llama-playground

KANNADA-LLAMA:
https://tensoic.streamlit.app/

OPENHATHI:
https://www.youtube.com/watch?v=WKfVzJSDAd8

KRUTRIM:
https://www.youtube.com/watch?v=5BhN0Qopt_0

BHASHINI:
https://www.youtube.com/watch?v=EBWBIZIE-VY

REQUIREMENTS FOR INDIC LLMS

- Data
  - digitization
  - tokenization
  - translation
  - curation
  - quality check
- Models
  - use pre-trained open-source models
  - train from scratch
  - fine-tuning for custom use cases
  - evaluation of model performance
  - distributed training on GPUs
- Deployment
  - inference speed and throughput
  - device: cloud, mobile, edge
- User Experience
  - localized to regional language
  - multi-modal experience
  - chat or voice
- Customers
- B2C - individual users
  - build for scale
  - personalization
  - self-learning
- B2B - enterprises
  - customizable
  - platform approach
  - train on own data

TAMIL-LLAMA: INTRODUCTION

- Developed by Abhinand, independent researcher and open-sourced paper, code and demo as below:
- https://huggingface.co/spaces/abhinand/tamil-llama-playground
- https://github.com/abhinand5/tamil-llama
- https://arxiv.org/abs/2311.05845

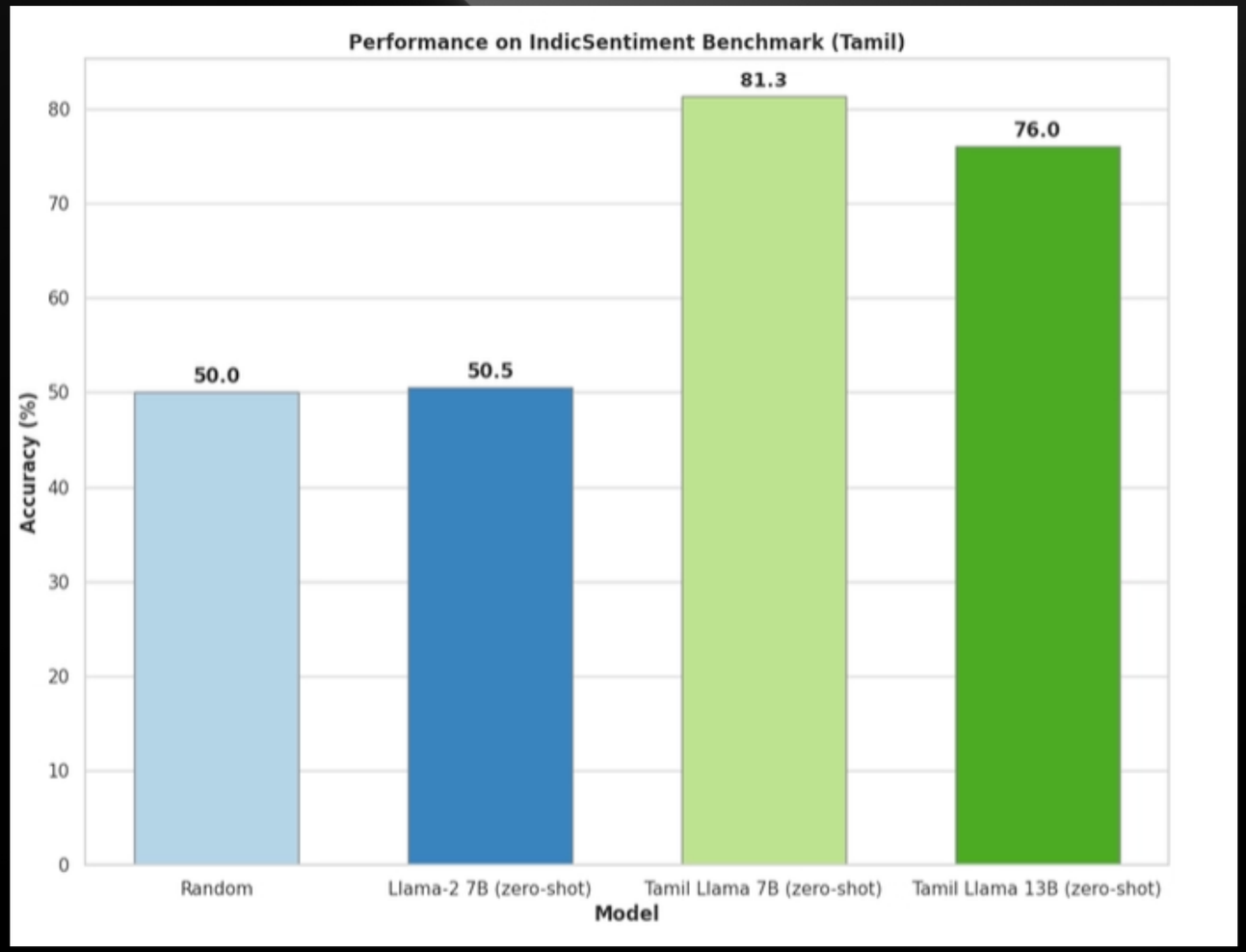
- Built on top of LLaMA-2 (Touvron et al., 2023) and released versions with 7B and 13B parameters
- Extended LLaMA's vocabulary from 32k to 48k tokens with 16k additional Tamil tokens (cf. Cui et al. 2023)
  - Used SentencePiece to train a Tamil Tokenizer
  - Integrated LLaMA-2's tokenizer with vocabulary from newly trained Tamil Tokenizer
  - Model can handle 3x more information and work 3x faster using fewer tokens
- Fine-tuning with LoRA to enable seamless UX through natural language queries with 145k instructions
- Release of Tamil-translated version of dataset (Alpaca)
- Marked improvement in performance of Tamil-LLaMA vs. GPT

- v0.2 released yesterday with bilingual capabilities in English and Tamil

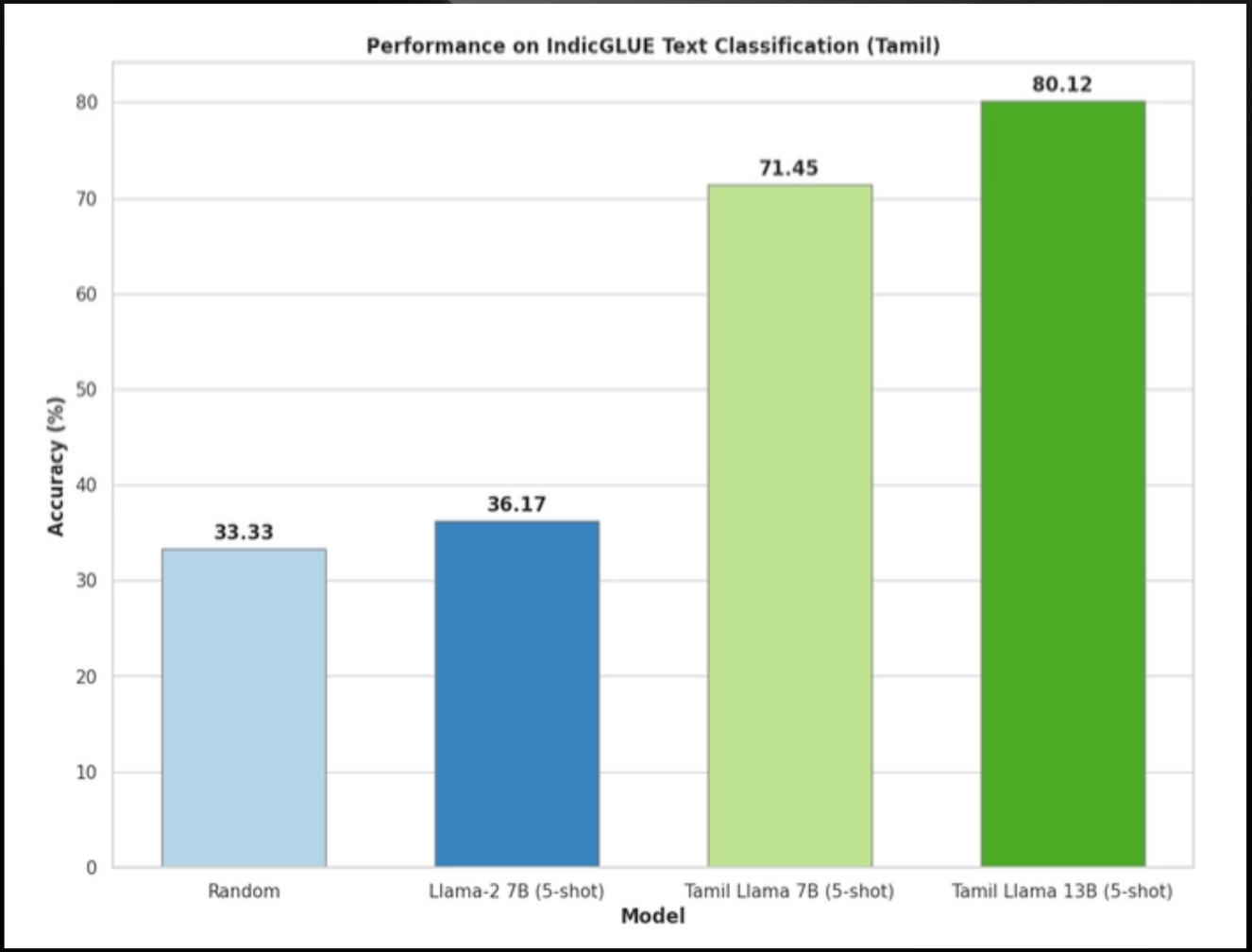
Table 3: GPT-4 rated performance scores for different models on Tamil instructions

Task Type	**Tamil-LLaMA-7B**	**Tamil-LLaMA-13B**	***gpt-3.5-turbo***
Question Answering	**77.00**	75.33	54.33
Open-ended QA	84.47	**85.26**	58.68
Reasoning	47.50	**64.25**	63.50
Literature	45.50	40.00	**71.00**
Entertainment	43.33	50.00	**60.00**
Creative Writing	92.50	**95.62**	59.69
Translation	60.56	66.67	**92.78**
Coding	63.57	**76.07**	57.14
Ethics	23.75	**57.50**	40.00
Overall	63.83	**71.17**	61.33

TAMIL-LLAMA: RESULTS



TAMIL-LLAMA: RESULTS

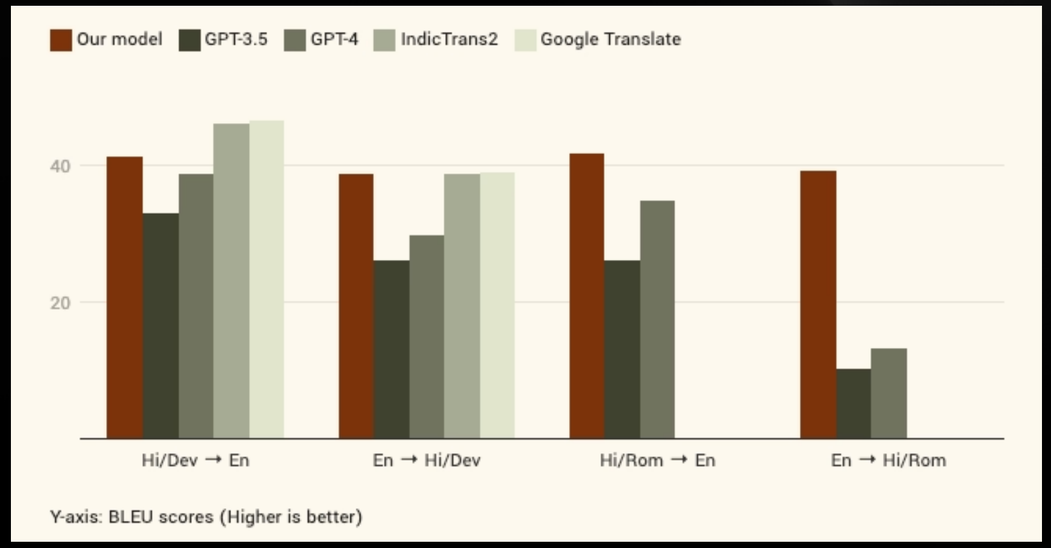


TAMIL-LLAMA: CHALLENGES

- **Constrained Knowledge Base**: Due to computational and cost constraints, our models were trained on a relatively limited Tamil dataset. This translates to gaps in the models' knowledge, especially regarding nuances and specifics native to Tamil culture and literature. While the current version lays the foundation, the true potential can be unlocked with access to a broader data spectrum, enriching its contextual understanding.
- **Ethical Concerns**: Detoxification procedures were not implemented in our training process, making these models prone to generating potentially harmful or offensive content. Their uncensored nature necessitates caution during deployment.
- **Lack of Robustness**: Our models may, at times, produce outputs that veer off-topic or deviate substantially from anticipated responses. This vulnerability is more pronounced under adversarial conditions or tricky prompts.
- **Reasoning and Mathematical Challenges**: While our models showcase competence in specific reasoning scenarios, they falter in many others, underscoring the repercussions of not having a comprehensive training set.
- **Over-Generation Tendencies**: On occasions, the models tend to generate verbose content, extending beyond logical termination points, leading to potential redundancy.
- **Evaluation Hurdles**: Assessment of LLMs is a crucial yet challenging endeavor. The scarcity of standardized benchmarks, particularly for languages like Tamil, which are outside the European linguistic group, complicates comparative evaluations. Although we propose an evaluative approach tailored for Tamil within this paper, it is not exhaustive enough to gauge models' efficacy across diverse domains.
- **Translation Loss**: Given that the instructional prompts used for fine-tuning the Tamil LLaMA base models are derived from English datasets translated into Tamil, there is a potential for nuanced inaccuracies—commonly referred to as translation loss. This can potentially affect the models' abilities in both text generation and comprehension due to subtle shifts in meaning that can occur during the translation process.

OPENHATHI (HINDI, HINGLISH)

- Developed by sarvam.ai (https://www.sarvam.ai/blog/announcing-openhathi-series)
- Extended LLaMA-2 Tokenizer with 16k additional tokens, to expand vocabulary size to 48k tokens
- Pre-training phase
  - Train the model for translation of text between Hindi & English & predict original text given its translation
  - Bilingual next token prediction task to increase alignment between Hindi and English
  - Fine-tuning on different datasets and benchmarks e.g. translation, toxicity classification etc.



OPENHATHI (HINDI, HINGLISH)

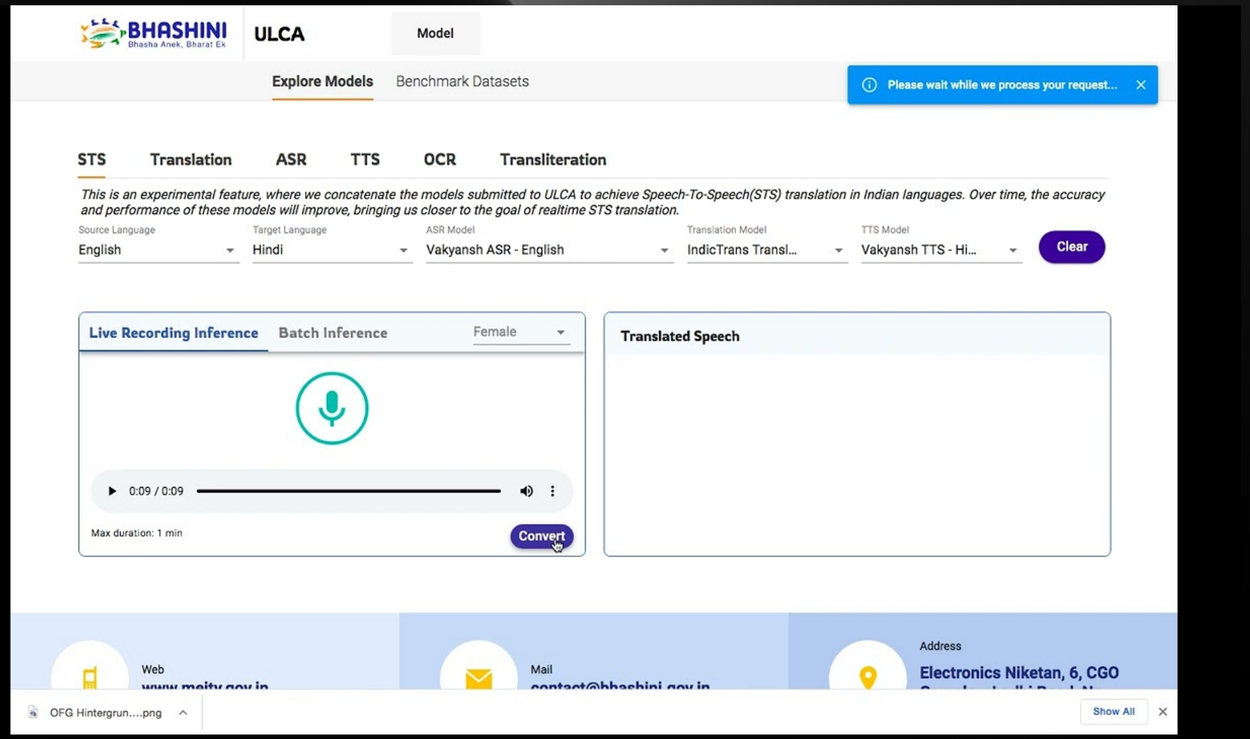
sarvam.ai

OpenHathi

KRUTRIM AI (MULTILINGUAL)

- Claim to have built a foundational model trained on over 2T tokens with data from multiple Indian languages
- Performance said to outperform LLaMA-2 on a variety of benchmarks
- LLM text generation capabilities in English, Hindi and 8 other languages; comprehension support for 22 languages
- Potential to combine voice to build accessible GenAI apps
- Also building multi-modal app combining text, vision and voice
- Technical details not available / shared with the community; waitlist for registration

- Language translation platform developed by MEITY (https://bhashini.gov.in/)



KANNADA LLAMA

Kannada LLM
https://www.tensoic.com/blog/kannada-llama/
https://tensoic.streamlit.app/

Kannada-English bilingual LLM
https://www.cognitivelab.in/blog/introducing-ambari

Followed similar strategy for vocabulary extension of LLaMA-2 vocabulary with additional Kannada tokens & training to Tamil-LLaMA and OpenHathi

OTHERS

- Telugu LLaMA (independent)
- Malayalam LLaMA (independent)
- Bharat GPT (corover.ai)
- Project Indus (Tech Mahindra)
- Bharat GPT (Reliance Jio)

WHAT NEXT?

- Leverage latest, state-of-the-art open-source foundational models
  - e.g. Mixtral vs. LLaMA-2

- Customization of LLMs
  - domain-specific use cases
    - e.g. law, governance, education, healthcare
    - e.g. KissanAI's Dhenu for agriculture - https://www.youtube.com/watch?v=Z-hXubdVTQ0

  - language-specific and multilingual LLMs
    - work in progress

- Multimodal LLMs
  - combine image and text understanding

- Voice UI/UX
  - connect voice input and output for seamless UX

OUTLINE

1 Large Language Models

2 LLMs for India

3 Business Use Cases

4 Careers & Resources

5 Q & A

Ongoing Efforts:

- **Government Initiatives:** The Indian government, recognizing the potential of LLMs, has launched initiatives like the National AI Mission to support research and development in this area.
- **Startup Ecosystem:** Several Indian startups are actively involved in building and deploying LLMs for various applications, focusing on vernacular languages and addressing specific Indian needs.

Challenges and Opportunities:

Despite the progress, building robust and representative LLMs for India faces challenges like:

- **Data Availability:** Access to high-quality, diverse, and ethically sourced data in Indian languages is crucial for training effective models.
- **Computational Resources:** Training large language models requires significant computing power, which can be a bottleneck for Indian researchers and startups.
- **Bias and Fairness:** Ensuring that LLMs trained on Indian data are free from bias and accurately reflect the country's cultural and linguistic diversity is critical.

USE CASES

- Customer service
- Chat bots
- Personalized user experience
- Content creation & marketing
- Employee productivity
- Enterprise Knowledge Management
- Document understanding
- Vernacular language interfaces
- Research & Analysis
- Content moderation
- Translation
- Summarization
- Translation: text-text & voice-voice etc.
- Agents

EXAMPLE USE CASES

EXAMPLE USE CASES

OUTLINE

1 Large Language Models

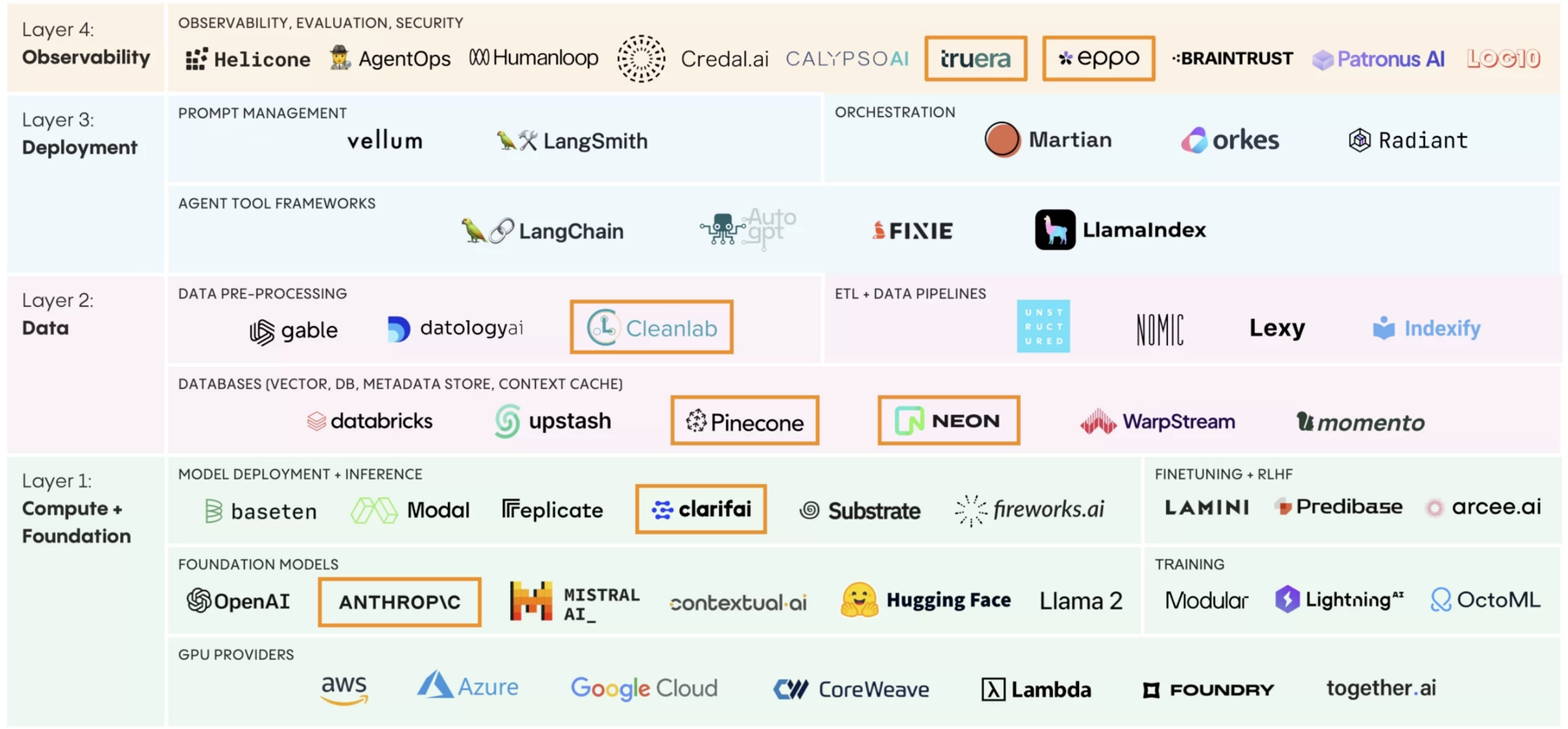
2 LLMs for India

3 Business Use Cases

4 Careers & Resources

5 Q & A

Modern AI Stack: The Emerging Building Blocks for GenAI



Backed by Menlo Ventures

https://menlovc.com/perspective/the-modern-ai-stack-design-principles-for-the-future-of-enterprise-ai-architectures/

COURSES

DEEPLEARNING.AI

- LLMOps
- ChatGPT Prompt Engineering for Devs
- Building Systems with the ChatGPT API
- LangChain for LLM App Development
- LangChain: Chat with your Data
- Building LLM Apps with LangChain.js
- Fine-tuning LLMs
- LLMs with Semantic Search
- Building GenAI Apps with Gradio
- Evaluating & Debugging GenAI Models
- Advanced Retrieval for AI
- Building and Evaluating RAG Apps
- Quality & Safety for LLM Apps
- Vector Databases
- Pair Programming with an LLM
- LLM Functions, Tools & Agents

OTHERS

- LLM Course by Maxime Labonne, JP Morgan - https://github.com/mlabonne/llm-course
- LLM Bootcamp by Full Stack Deep Learning - https://fullstackdeeplearning.com/llm-bootcamp/spring-2023/
- Practical courses and certifications from Cloud companies e.g. Google Cloud - https://cloud.google.com/blog/topics/training-certifications/new-generative-ai-trainings-from-google-cloud

LLM CAREER ADVICE

Develop a life-long learning mindset

Stay abreast of the latest developments about LLMs & GenAI and be eager and willing to learn new tools

Build domain expertise

For GenAI use cases in fields of interest e.g. e-commerce, finance, education, healthcare

All software developers are now AI developers

New GenAI roles e.g. LLM Engineer, Prompt Engineer apart from transforming the AI/ML Engineer/Scientist roles

Deciding which company to join to build GenAI

Optimize for roles at companies where GenAI/LLM use cases is critical for the business vs. a marketing stunt

Share, communicate and network

Meetups to interact with founders, engineers, VCs etc. to understand real-world use cases and opportunities

Long-term AI career planning

Plan your career at various time scales: short-term (1-3 years) and longer term (5+ years); be open to change & evolve with the field

DECIDE ON THE RIGHT FIT/ROLE

- LLM Engineer
- Prompt Engineer
- AI/ML/Research Engineer
- AI/Research Scientist
- Software Engineer - ML

BUILD YOUR GENAI PORTFOLIO

- Learn by building
- Build end-to-end apps
- Add portfolio projects to your CV
- Share your work on LinkedIn, X etc.
- Present your work in meetups, webinars

RESOURCES

TAMIL, TELUGU & MALAYALAM

https://abhinand05.medium.com/breaking-language-barriers-introducing-tamil-llama-v0-2-and-its-expansion-to-telugu-and-malayalam-deb5d23e9264

https://github.com/abhinand5/tamil-llama

https://huggingface.co/spaces/abhinand/tamil-llama-playground

https://www.medianama.com/2023/12/223-folklore-ai-swecha-telugu-chatbot/

HINDI & MULTILINGUAL

https://corover.ai/bharatgpt/

https://www.sarvam.ai/blog/announcing-openhathi-series

https://olakrutrim.com/

https://www.techmahindra.com/en-in/innovation/the-indus-project/

KANNADA

https://www.tensoic.com/blog/kannada-llama/

https://tensoic.streamlit.app/

https://www.cognitivelab.in/blog/introducing-ambari

https://analyticsindiamag.com/tensoic-releases-playground-for-kannada-llama-on-nvidia-a100s/

OUTLINE

1 Large Language Models

2 LLMs for India

3 Business Use Cases

4 Careers & Resources

5 Q & A

QUESTIONS

You can find me on:

Slides available at sundeepteki.org/talks

Pitch

Want to make a presentation like this one?

Start with a fully customizable template, create a beautiful deck in minutes, then easily share it with anyone.

Create a presentation (It's free)