4. Large Language Models- Science & Technology

How are Indian firms training LLMs?. India's entry into the LLM space with Sarvam AI's 105B model signals a move toward sovereign AI. By utilizing the IndiaAI Mission's GPU clusters and adopting Mixture of Experts (MoE) architectures, Indian firms are working to bypass the "English bias" and the high "tokenization costs" of global models. Success depends on building a robust Indian language dataset and focusing on domain-specific models for local governance and healthcare.

1. Understanding Large Language Models (LLMs)

Definition- Advanced AI systems designed to process, understand, and generate human-like text by learning patterns from massive datasets.

Architecture- Most modern LLMs use Deep Learning with Transformer Architectures (like GPT). These architectures utilize an "attention mechanism" that allows the model to weigh the importance of different words in a sentence to understand context.

Mechanism- Words are converted into tokens, which are then transformed into numerical embeddings. These embeddings represent the semantic and conceptual relationships between words in a multidimensional space.

2. Training Techniques and Enhancements

LLMs go through a rigorous training cycle to ensure they can generalize across tasks-

Self-Supervised Learning- The model learns by predicting the next word in a sentence from a massive text corpus without explicit human labelling.

Zero-Shot Learning- The ability of a model to perform a task or make predictions about data it was not specifically exposed to during training.

RLHF (Reinforcement Learning with Human Feedback)- A crucial step where human trainers rank model responses to minimize biases, improve safety, and enhance accuracy.

Prompt Engineering- The process of refining inputs to guide the model toward more accurate and relevant outputs.

3. Comparative Analysis- Standard LLM vs. Mixture of Experts (MoE)

To overcome high costs, newer models like those from Sarvam AI often look toward **MoE** architectures.

Feature	**Standard LLM**	**Mixture of Experts (MoE)**
Activation	All parts of the model work for every single query.	Only specialized sub-parts (experts) are activated per query.
Resource Use	High power and compute consumption.	Significantly faster and more cost-effective.
Analogy	Every teacher in a school answer every question.	Only the math teacher answers a math question.
Efficiency	Lower efficiency for massive parameter counts.	High efficiency; allows for more parameters with less "active" compute.

4. Challenges for Indian LLM Development

Developing indigenous LLMs faces unique structural and financial hurdles-

Prohibitive Capital Intensity- Training a 70-billion-parameter model costs approx. $6 million.

1. GPT-4 required tens of millions of dollars and thousands of GPUs running for months.

Linguistic Data Scarcity- The web is dominated by English (>50%), while most Indian languages account for less than 1% each.

1. Common Crawl and other global datasets have minimal representation of Hindi, Tamil, or Bengali.

The "Tokenization" Penalty- Models often translate Indian languages to English and back, increasing inference costs.

1. A 10-word sentence uses ~15 tokens in English but 20–25 tokens in Hindi due to inefficiencies in script processing.

5. Government Support and Indigenous Efforts

The Indian government has recognized AI as a strategic asset through the IndiaAI Mission-

GPU Subsidies- The IndiaAI Mission commissioned 36,000 GPUs (via providers like Yotta) to provide affordable compute access.

Direct Support to Sarvam AI- The government allocated 4,096 GPUs with subsidies valued at nearly ₹100 crore to help Sarvam train its 105B parameter model.

Institutional Players- * BharatGen (IIT Bombay)- Multilingual 17B parameter model focused on education and healthcare.

1. Gnani.ai- Focused on speech-based AI and text-to-speech models.

6. Way Forward for India's AI Ecosystem

1. Bhashini Initiative- Using public-private partnerships to create high-quality annotated corpora in regional languages.
2. Sector-Specific Models- Instead of building "general" models, focusing on smaller, highly efficient models for Agriculture, Law, and Governance.
3. Energy Efficiency- Adopting MoE and Model Compression techniques to reduce the massive electricity and compute footprint.
4. Academic Integration- Strengthening the link between IITs/IIITs and startups to ensure a steady supply of researchers skilled in Large Language Model training.

Source- https-//www.thehindu.com/sci-tech/technology/how-are-indian-firms-training-llms-explained/article70676898.ece