BUILDING

LARGE LANGUAGE MODELS

for Local and Regional Applications

Center for AI & BigData

Namal University Mianwali

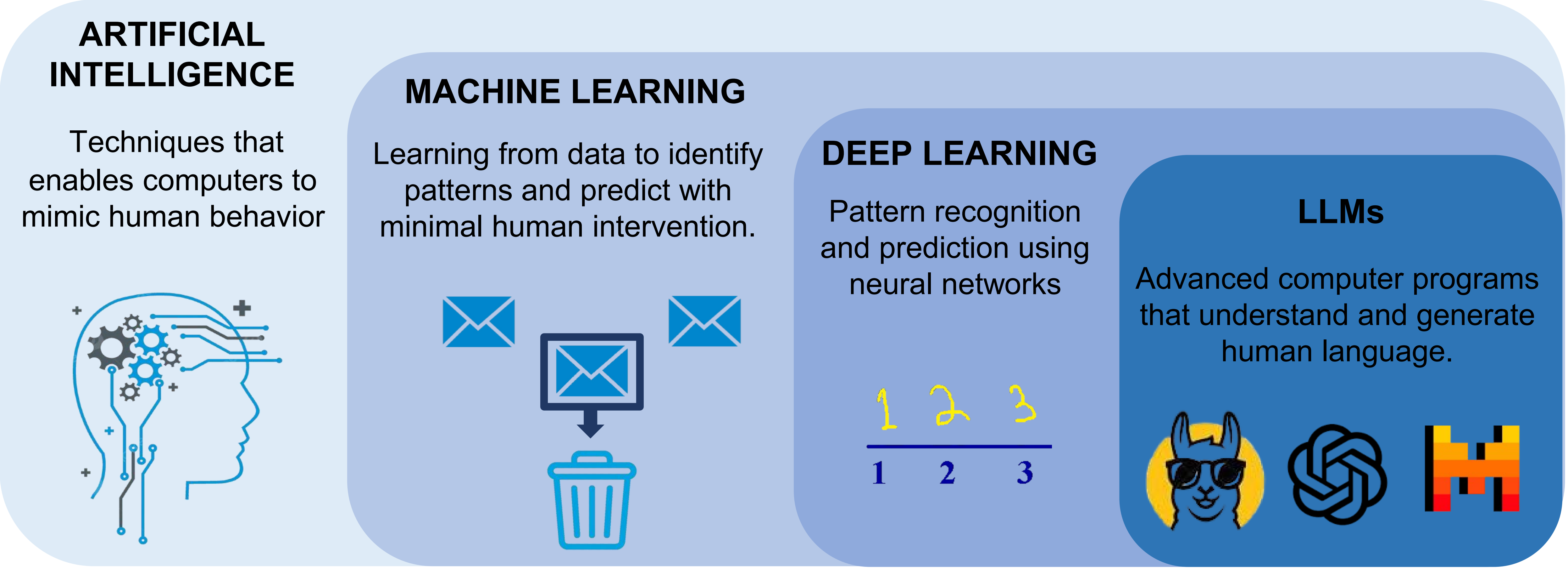
Presented by: Kamran Younis | Research Assistant

Contents

- A quick intro to LLMs
- Language Gap and Urdu/Regional Challenges
- LLMs Open-Source Ecosystem
- Building LLMs for Regional Challenges
- Local LLM Architectures
- Resource Optimization, Open-Source Fine-Tuning
- Applications & Use Cases
- Implementation Strategy

INTRODUCTION AND CONTEXT

What are LLMs?



ARTIFICIAL INTELLIGENCE

Techniques that enables computers to mimic human behavior

MACHINE LEARNING

Learning from data to identify patterns and predict with minimal human intervention.

DEEP LEARNING

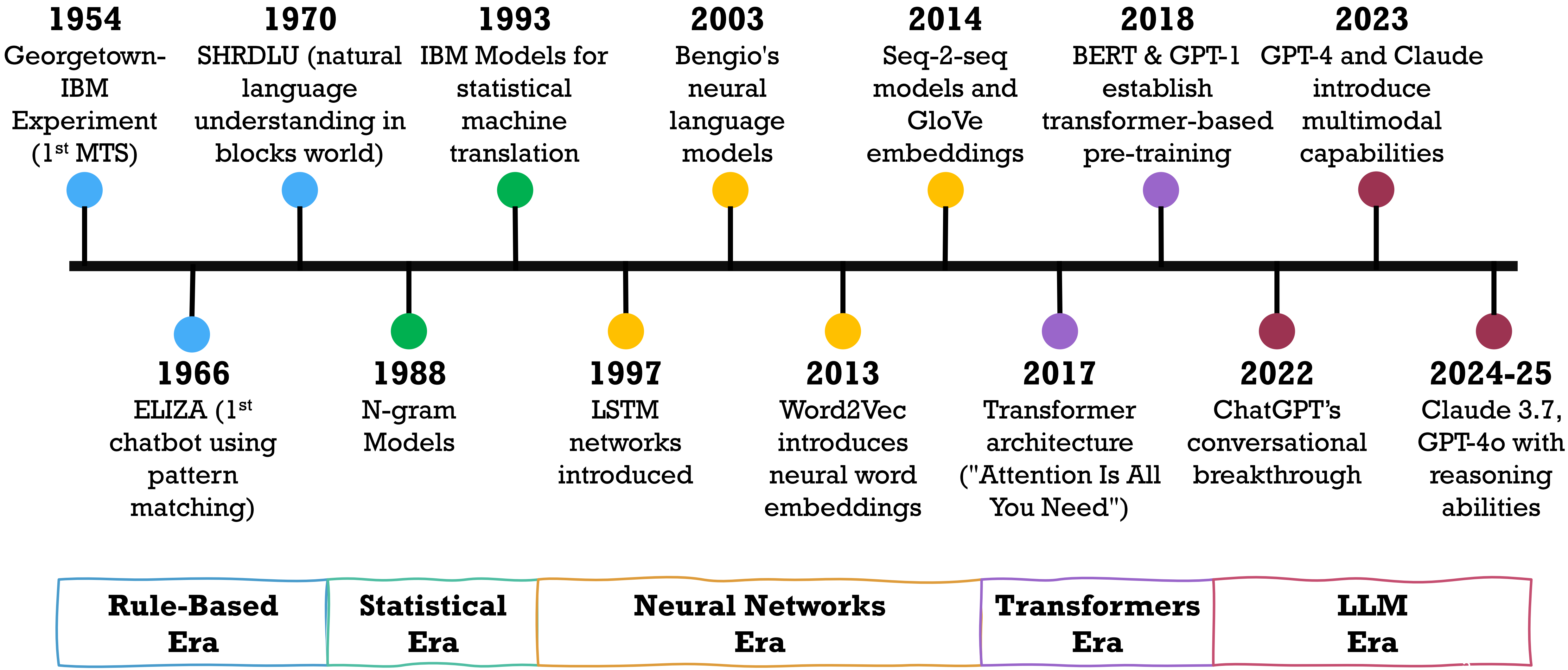
Pattern recognition and prediction using neural networks

LLMs

Advanced computer programs that understand and generate human language.

“A Large Language Model is an AI that processes and generates human language using vast text data and deep neural networks.

Evolution of Language Understanding



1954
Georgetown-IBM Experiment (1st MTS)

1966
ELIZA (1st chatbot using pattern matching)

1970
SHRDLU (natural language understanding in blocks world)

1988
N-gram Models

1993
IBM Models for statistical machine translation

1997
LSTM networks introduced

2003
Bengio's neural language models

2013
Word2Vec introduces neural word embeddings

2014
Seq-2-seq models and GloVe embeddings

2017
Transformer architecture ("Attention Is All You Need")

2018
BERT & GPT-1 establish transformer-based pre-training

2022
ChatGPT's conversational breakthrough

2023
GPT-4 and Claude introduce multimodal capabilities

2024-25
Claude 3.7, GPT-4o with reasoning abilities

Rule-Based Era

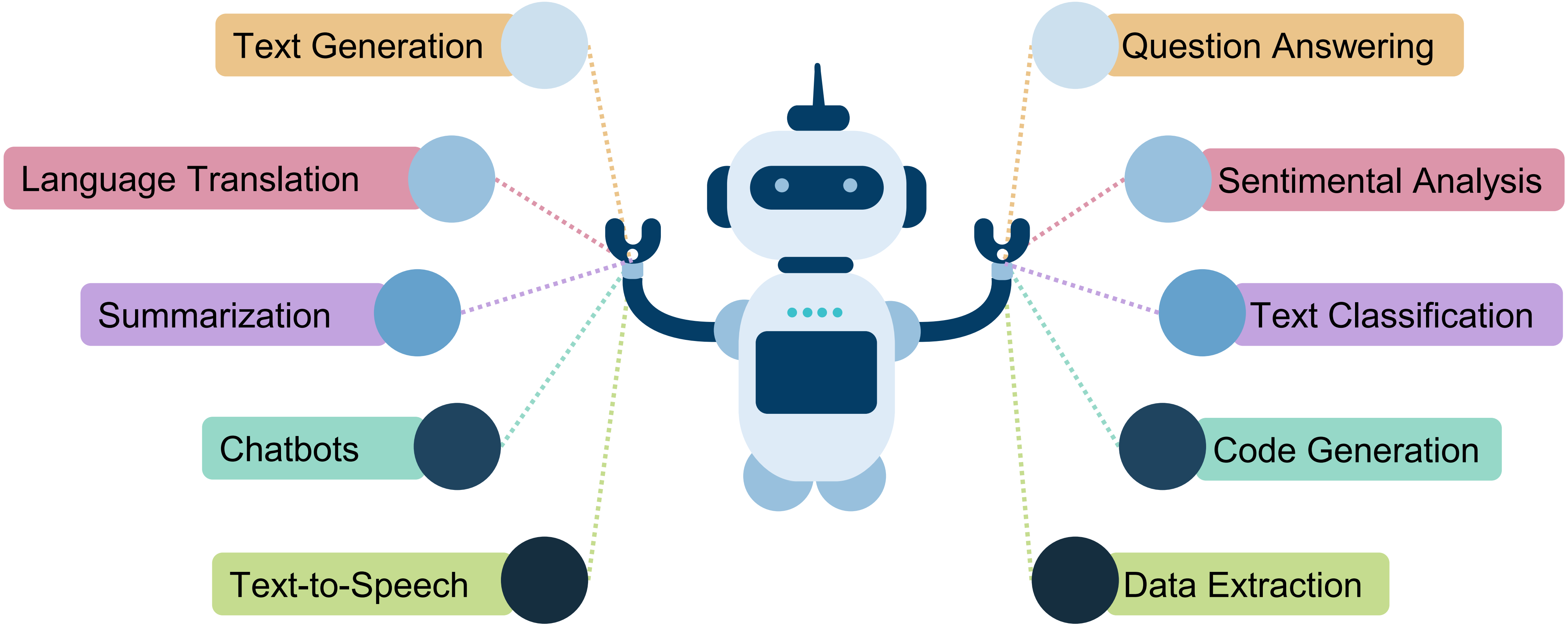
Statistical Era

Neural Networks Era

Transformers Era

LLM Era

What They Can Do?



Real-World Impact Across Industries

All Industries:
LLMs can now automate '60-70% of employees' time. [McKinsey]

Agriculture:
Decrease operating costs by 22% [ARK]

Consulting:
25.1% faster, 40% higher quality [Harvard/BCG]

Healthcare:
58.7% better interpersonal skills [Nature]

Software development:
55.8% faster [Microsoft/MIT]

Medicine:
Charts 75% faster, 250% more detail [Carbon Health]

Finance, accounting, auditing:
Fully exposed, 100% faster [OpenAI/UPenn]

Therapy/coaching:
AI promises more effective therapy [Seligman]

Chemistry:
GPT-4 represents the future of chemistry [White]

Customer service:
14% more issues resolved [McKinsey]

Writers:
Fully exposed, 100% faster [OpenAI/UPenn]

The Current AI Race (2024–2025)

❑ **Reasoning:**

- OpenAI o1/o3: "Thinking time" solves Math Olympiad problems (beating humans)
- DeepMind’s **Gemini Deep Think** hits **gold medal** IMO level

❑ **Multimodal → video next:**

- **Standard:** Text + vision (GPT-4V, Claude-3, Gemini)
- **Next:** High-quality video generation (Sora, Veo3, Runway)

❑ **Agentic Use:**

- Microsoft Copilot: hundreds of millions of Office users
- Claude used in IDEs like Replit/Cursor for code & automation

❑ **Real-World Tests:** Kaggle AI Chess tournament - OpenAI's o3 beat xAI's Grok-4

❑ **Open source catching up:** LLaMA 3.1, Qwen 2.5, Kimi K2, GLM 4.5 - Beating commercial AIs on key benchmarks

WHY REGIONAL LLMs?

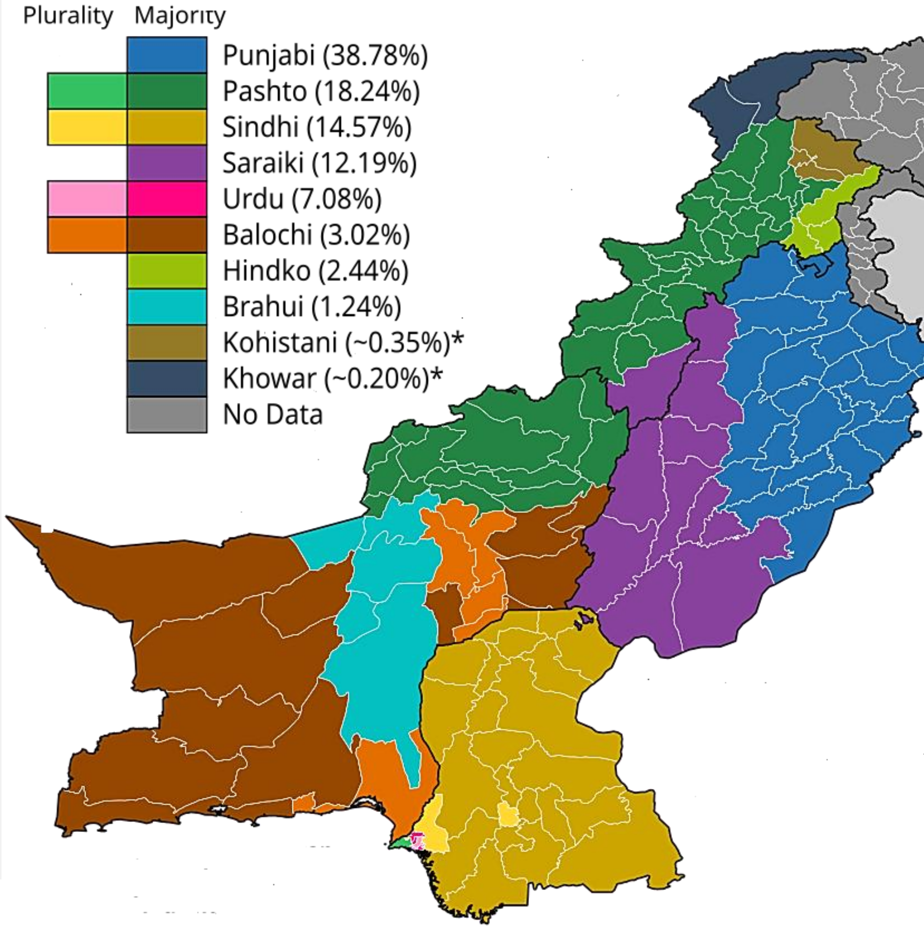
The Language Divide in Technology

The Global Challenge:

- 7,100+ languages spoken worldwide
- Top 20 languages represent 50% of global speakers
- 98% of AI research focuses on fewer than 10 languages
- English dominates despite being spoken by only 18% of the world population

The Regional Reality:

- Millions of people lack access to AI in their native languages
- Critical services, education, and information remain inaccessible

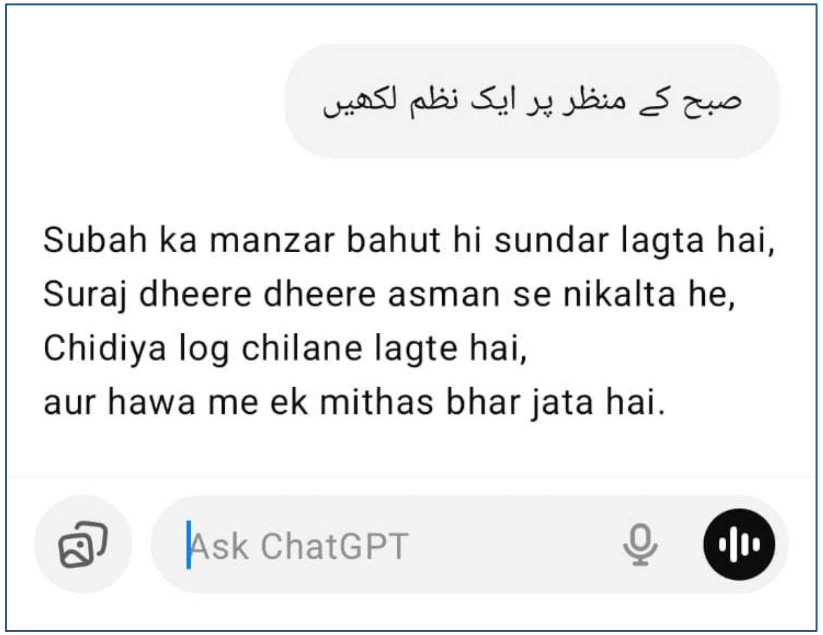


Pakistan's Linguistic Landscape as of the 2017
Source: Wikipedia (Pakistani census 2017)

The Language Gap - Technical Evidence

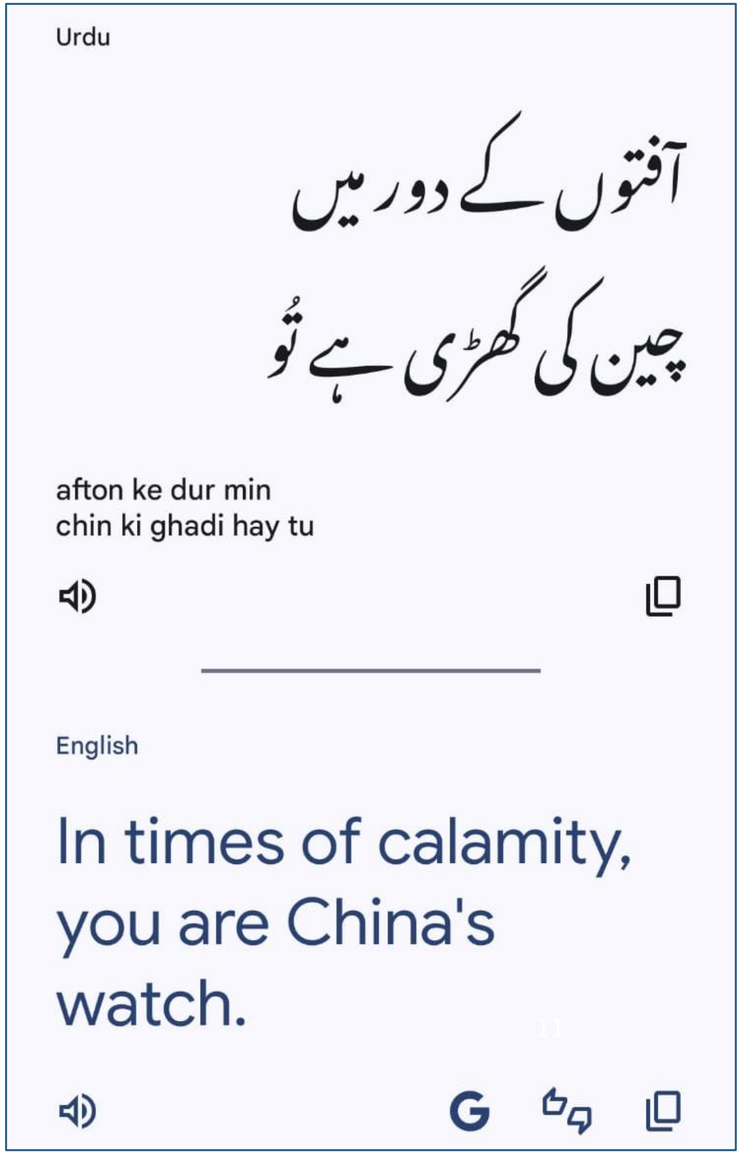
Translation vs. Understanding:

- Google Translate Urdu accuracy: 67% (compared to 94% for Spanish)
- ChatGPT cultural context accuracy in Urdu: 71%
- Urdu content in major AI training datasets: <0.05%



صبح کے منظر پر ایک نظم لکھیں

Subah ka manzar bahut hi sundar lagta hai,
Suraj dheere dheere asman se nikalta he,
Chidiya log chilane lagte hai,
aur hawa me ek mithas bhar jata hai.



Urdu

آفتوں کے دور میں
چین کی گھڑی ہے تُو

afton ke dur min
chin ki ghadi hay tu

English

In times of calamity, you are China's watch.

Challenges in Regional LLM Development

The Script Challenges

- **Right-to-left complexity:** Arabic script requires 40% more computational resources than Latin scripts
- **Contextual variations:** Single Urdu word can have 12+ different forms depending on position
- **Code-switching reality:** 73% of Pakistani digital users mix Roman-Urdu with English in conversations

Data Challenges

- **Training data scarcity:** Urdu represents <0.05% of major AI datasets vs. 54% English
- **Quality gap:** Only 2.3M high-quality Urdu sentences available vs. 570M for Spanish
- **Domain gaps:** Legal, medical, and technical Urdu content virtually non-existent in training data

The Solution - Local & Regional Language Models

What We Need: Pakistan-Native AI Systems

- **Native Processing:** Fluent in Urdu, Punjabi, Sindhi, and local English patterns.
- **Cultural Intelligence:** Deep understanding of Islamic values, Pakistani traditions, and social norms.
- **Local Expertise:** Knowledge of Pakistani laws, geography, history, and current affairs.
- **Contextual Awareness:** Understanding of regional business practices, educational systems, and daily life.

Agriculture
Business & Finance
Education
Govt. Services
Healthcare
E-Commerce

Existing Initiatives and Their Limitations

Zahanat AI

- **Pakistan's first indigenous GPT** launched by Data Vault (March 2025)
- **Scale**: 1.5B parameters based on Meta's LLaMA architecture
- **Development cost**: <$1M (vs. $5M for DeepSeek)
- **Current status**: Beta testing, limited access within Pakistan
- **Limitations**: English-dominant, basic Urdu support, limited cultural training

Jazz-NUST-NITB Collaboration (Nov 2024)

- **Scope:** 5-year MoU for indigenous Urdu LLM development
- **Languages:** Primary focus on Urdu, datasets for Pashto and Punjabi
- **Challenge:** Still in early development phase, no public release timeline

Existing Initiatives and Their Limitations

Alif 1.0 by Traversaal AI (Feb 2025)

- **Performance**: Outperforms Meta LLaMA 3.1 8B on Urdu-specific tasks
- **Innovation**: Native Urdu reasoning with cultural alignment
- **Status**: Open-source model available on HuggingFace

Still, we have limitations!

- **Market fragmentation:** 6+ separate initiatives with no coordination
- **Limited scale:** Largest model only 8B parameters vs. 175B+ international standards
- **Data scarcity:** <0.1% of internet content in Urdu affects all projects
- **Commercial viability:** No profitable business model demonstrated yet

Learning from Regional AI Success Stories

Indic Language Models (India)

- **AI4Bharat's IndicBERT**: Serves 600M users across 22 languages
- **Commercial success**: $12M funding, partnerships with Google, Microsoft
- **Impact**: 340% improvement in local language tasks vs. multilingual models

Arabic AI Initiatives (Middle East)

- **JAIS Model (UAE)**: $100M investment, 13B parameters trained on Arabic
- **Results**: 89% accuracy in Arabic vs. 34% for GPT-3.5
- **Market capture**: 67% of regional enterprise AI adoption

THE OPEN-SOURCE ECOSYSTEM

Commercial vs Open Source LLMs

Closed Commercial Leaders:

- ➢ **OpenAI GPT-4:** $20/month, API access only
- ➢ **Anthropic Claude:** API-only, usage limits
- ➢ **Google Gemini:** Integrated products, limited access

Open-Source Models:

- ✓ **DeepSeek V3/R1:** 671B parameters. 30x cheaper than OpenAI o1
- ✓ **Kimi K2:** 1T parameters. #1 agentic AI, 90.6% tool success rate
- ✓ **Qwen 3:** 235B parameters. 119 languages, beats DeepSeek-R1
- ✓ **GLM 4.5:** 355B parameters. #3 globally, hybrid reasoning
- ✓ **OpenAI gpt-oss:** First open weights since GPT-2 (finally!)

Open-Source Models are Taking the Lead!

Western Open-Source Model Families

LLaMA Family (Meta):

- ✓ **LLaMA 4** (April 2025): Scout & Maverick variants, 256K context, multimodal
- ✓ **LLaMA 3.3**: 70B parameters, 128K context window
- ✓ **Code Llama**: Programming specialist, GitHub integration

Legacy: Started the open-source revolution, 40M+ downloads

Mistral Models (Mistral AI):

- ✓ **Mixtral 8x22B**: Latest MoE, 141B total parameters
- ✓ **Mistral Large**: Competing with GPT-4 class models

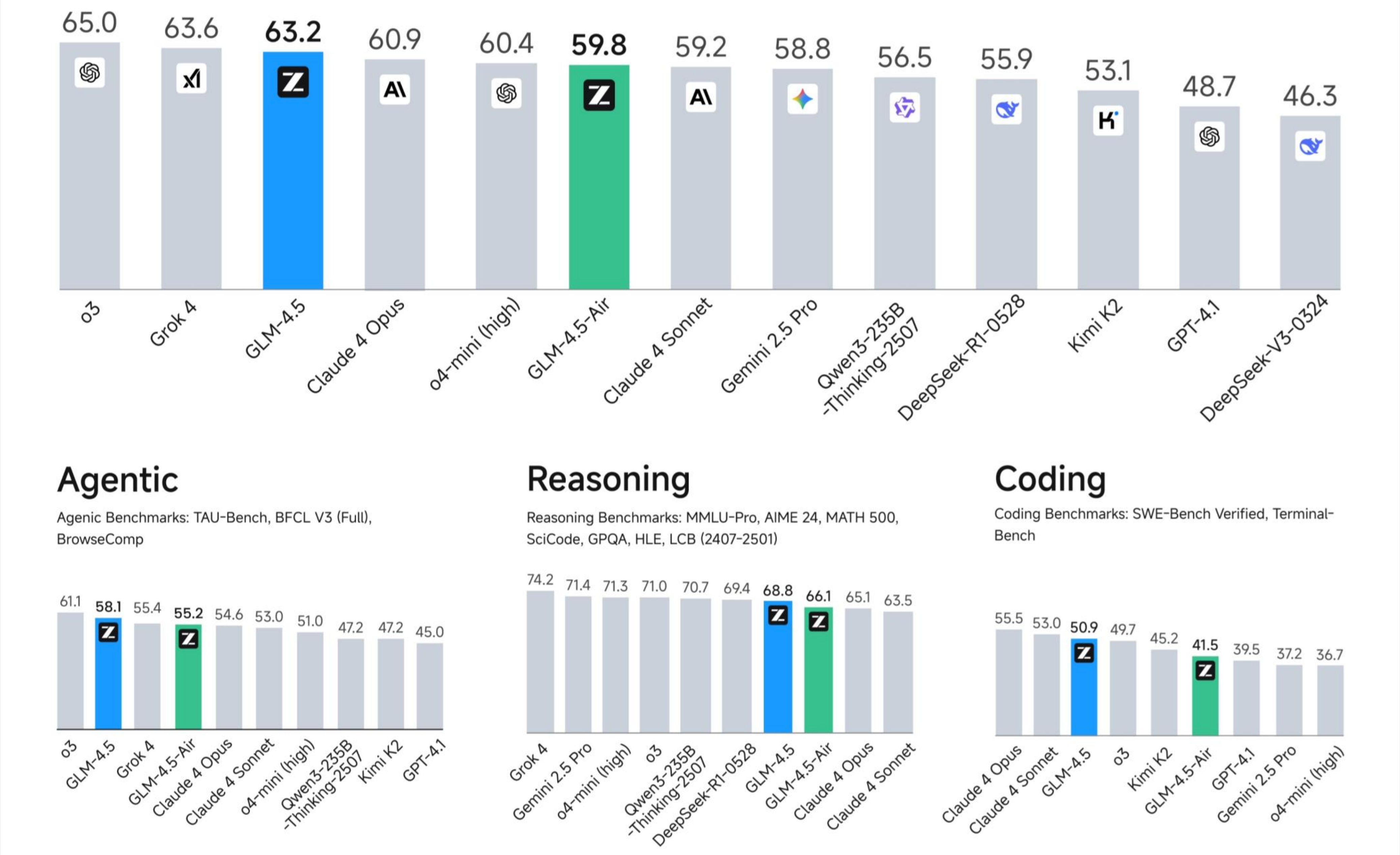
Chinese Dominance (2025 Breakthrough)

Family	Company	Model	Parameters	Features & Achievements
Qwen	Alibaba	Qwen 3	235B MoE	119 languages, Beats DeepSeek-R1
		Qwen 2.5-Max	-	Outperforms DeepSeek V3 in security
DeepSeek	-	DeepSeek V3	671B MoE	Industry-leading cost efficiency
		DeepSeek R1	-	Reasoning specialist, 30x cheaper than o1
Kimi	Moonshot AI	Kimi K2	1T	Agentic AI leader, 90.6% tool success rate
GLM	Z.ai	GLM 4.5	355B	Hybrid thinking modes, #3 globally ranked
		GLM 4.5-Air	106B	Efficient for edge deployment

OpenAI gpt-oss (August 2025): 120B & 20B variants, Apache 2.0 license

Significance: Big Tech finally joins open-source movement!

Chinese Dominance (2025 Breakthrough)



Agentic

Agenic Benchmarks: TAU-Bench, BFCL V3 (Full), BrowseComp

Reasoning

Reasoning Benchmarks: MMLU-Pro, AIME 24, MATH 500, SciCode, GPQA, HLE, LCB (2407-2501)

Coding

Coding Benchmarks: SWE-Bench Verified, Terminal-Bench

The Open-Source Ecosystem

Hugging Face: 500K+ models, 100K+ datasets, 1M+ developers, 10B+ downloads

Model Serving & Deployment

□ **vLLM:** High-throughput inference server, 24x faster than vanilla PyTorch

□ **Ollama:** Local model management with one-command deployment

□ **TensorRT-LLM:** NVIDIA optimized serving for hardware acceleration

□ **LM Studio:** User-friendly interface for no-code local AI deployment

Training Frameworks:

✓ **Transformers (Hugging Face):** Industry standard with 180K+ GitHub stars

✓ **DeepSpeed (Microsoft):** Distributed training for trillion-parameter models

✓ **Accelerate:** Simplified multi-GPU training framework.

BUILDING REGIONAL MODELS

Broader Vision

Phase 1: R&D

- **Data Collection**: Build comprehensive Urdu corpus
- **Model Development**: Complete initial training experiments
- **Baseline Testing**: Establish performance metrics against existing solutions

Phase 2: Testing

- **Domain Applications**: Test in education, healthcare, agriculture
- **User Feedback**: Deploy with limited user groups
- **Model Refinement**: Iterate based on real-world performance

Phase 3: Market Entry

- **Production Deployment**: Launch stable commercial version
- **Partnership Development**: Establish B2B and government relationships
- **Scale Planning**: Prepare for broader market adoption

Building the Data Foundation

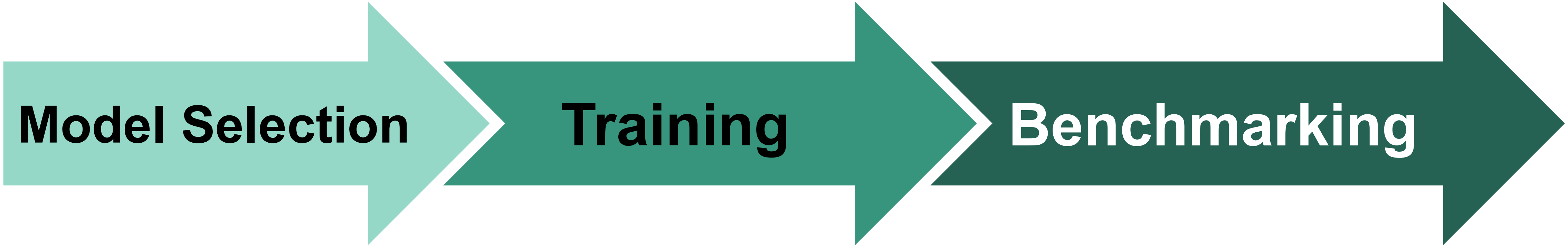
The Current Reality

- **Digital Urdu content available**: ~15TB across all sources
- **International LLM standard**: 500TB+ high-quality text required
- **Data collection costs**: PKR 20 per 1K tokens vs. PKR 0.80 for English content

Strategy

- ✓ Literary Corpus: Digitize 847 classic and modern Urdu books
- ✓ News Media: Collect 1.2M articles from 15 major publications (2018-2024)
- ✓ Social Media: 50M posts with cultural context annotations
- ✓ Academic Content: Gather 180K research papers from Pakistani universities
- ✓ Religious sensitivity: Develop Islamic jurisprudence knowledge base
- ✓ Social understanding: Map family structures and respect hierarchies
- ✓ Regional diversity: Include Punjabi, Sindhi, Balochi cultural contexts

Technical Approach for Regional LLMs



Model Selection

Training

Benchmarking

- Open source LLaMA 2-7B Models
- Mistral Models
- Custom tokenizer

- Continual Pre-training
- Cultural Fine-tuning
- Multi-stage approach

Urdu comprehension: Aim to exceed current GPT-4 performance on Urdu tasks

Code-switching: Handle mixed Urdu-English communication patterns

Training Large Language Models

Three Stage Pipeline

Stage	Objective	Outcome
Pre-training	Learn language from internet	General intelligence
Supervised Fine-tuning	Learn to follow instructions	Conversational ability
RLHF	Align with human values	Safe, helpful responses

Stage 1: Pretraining

The Data Diet

- ✓ Urdu News media, web pages, books, wikipedia, literary content, internet and social media etc.
- ✓ Max no. of tokens of human Urdu text

Learning Progression

Phase 1: Language structure and syntax

Phase 2: Factual knowledge acquisition

Phase 3: Reasoning capabilities

Computational Requirements

- ➢ Multi GPU cluster deployment
- ➢ Continuous training for 3-6 months

Stage 2: Supervised Fine Tuning

Supervised Fine-tuning (SFT):

Goal: Make model follow specific instructions

How: Feed curated human-written local language examples (input → correct output)

Model starts producing task-relevant responses after SFT

Why SFT matters:

Pre-training makes a **knowledgeable model**, but not necessarily an **obedient assistant**.

SFT shapes *what the model does with its knowledge*.

Stage 3: Reinforced Learning with Human Feedback

Process:

Collect preferences – Experts rank different model outputs

Train reward model – learns what “good” means

Policy optimization – model adapts to produce preferred answers

Outcome:

✓ Safer, more helpful, less biased responses

✓ Behaves closer to expectations

Training Infrastructure

Parallelism Strategies:

- **Data Parallelism:** Replicate model per GPU → train on different batches → sync gradients.
- **Model Parallelism:** Split model layers across GPUs → share parameters via communication.
- **Pipeline Parallelism:** Assign different layers to GPUs → process multiple micro-batches at once.
- **Tensor Parallelism:** Split single ops (e.g., attention) across GPUs → high communication cost.

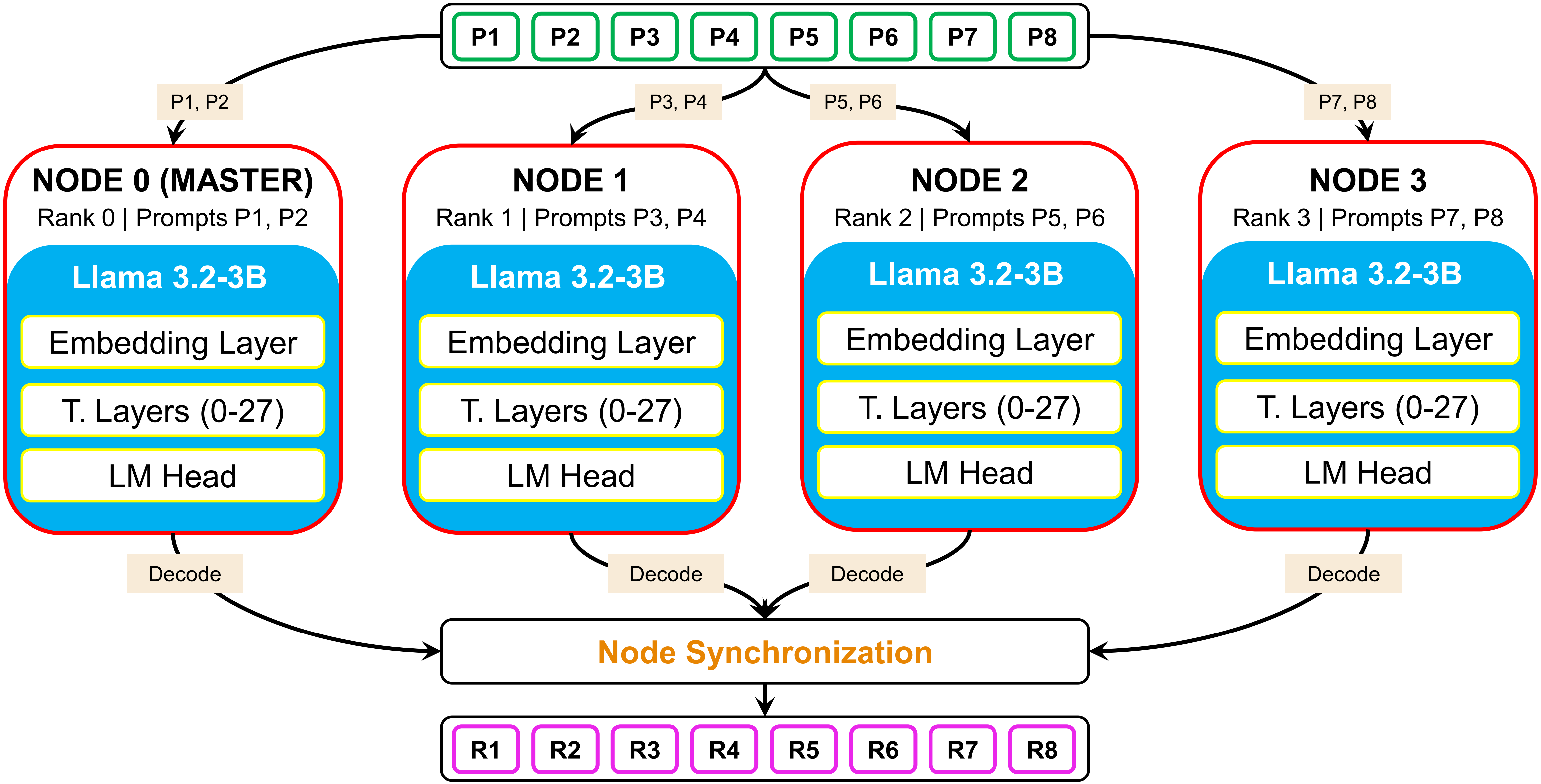
Example – Namal HPC Cluster:

✓ 6-node GPU cluster with SLURM job management

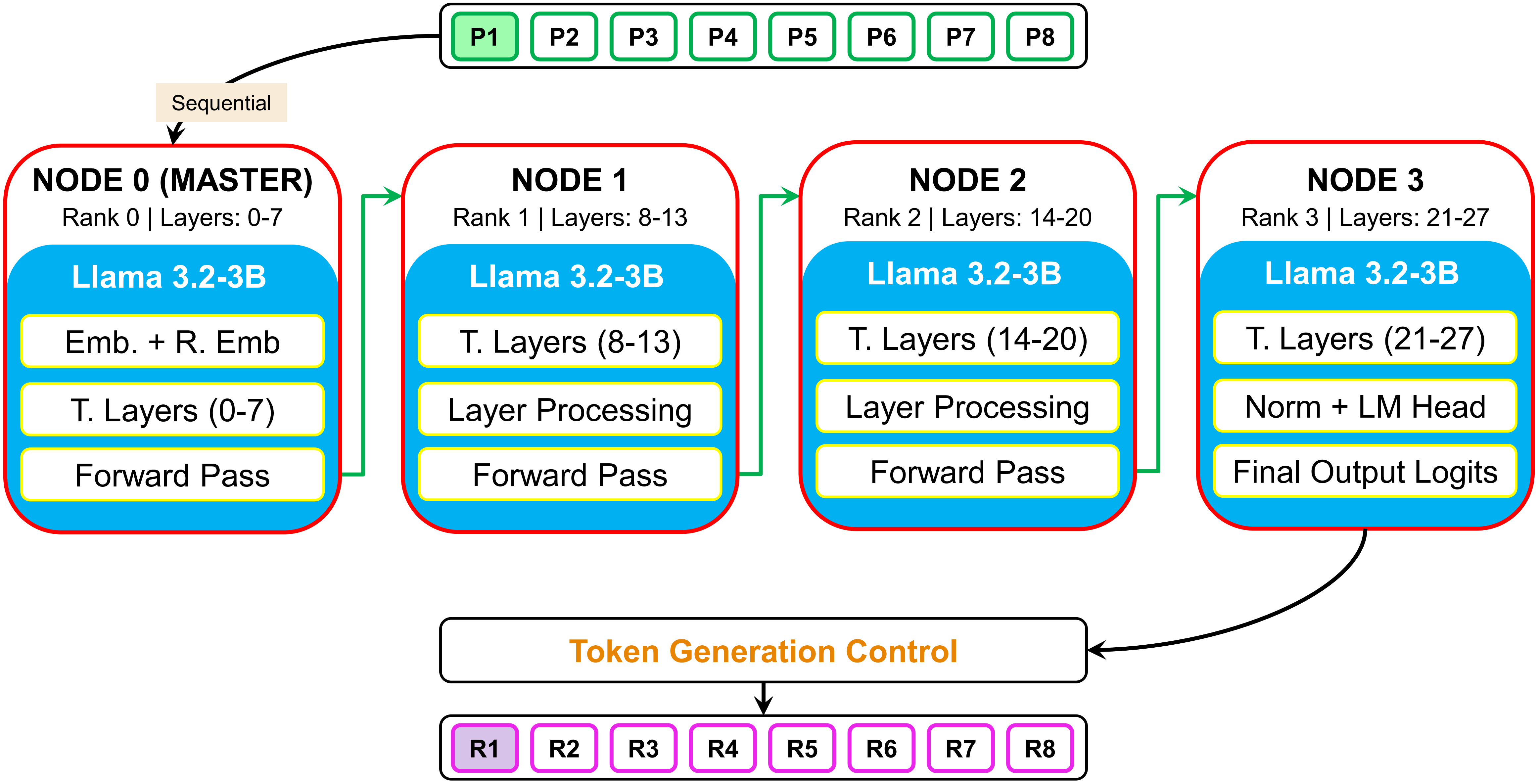
✓ GPUs: RTX 4070 Ti (12GB VRAM each) = 72GB total GPU memory

✓ Successfully implemented distributed training and inference of open source LLMs.

Data Parallelism



Model Parallelism



P1 P2 P3 P4 P5 P6 P7 P8

Sequential

NODE 0 (MASTER)
Rank 0 | Layers: 0-7
Llama 3.2-3B
Emb. + R. Emb
T. Layers (0-7)
Forward Pass

NODE 1
Rank 1 | Layers: 8-13
Llama 3.2-3B
T. Layers (8-13)
Layer Processing
Forward Pass

NODE 2
Rank 2 | Layers: 14-20
Llama 3.2-3B
T. Layers (14-20)
Layer Processing
Forward Pass

NODE 3
Rank 3 | Layers: 21-27
Llama 3.2-3B
T. Layers (21-27)
Norm + LM Head
Final Output Logits

Token Generation Control

R1 R2 R3 R4 R5 R6 R7 R8

Hardware Execution Pipeline & Latency Sources

Model Loading → Tokenization → Layer Processing → Token Generation → Output Decode

Cold Start: NVMe → RAM → GPU (2-5s), CUDA initialization (200-500ms)

Input Processing: Tokenization (5-10ms), embedding lookup (1-3ms)

Core Computation: Attention $O(n^2)$ operations (50-200ms/layer)

Token Generation: Sampling/softmax (2-5ms/token), KV cache updates (1-3ms/token)

Output: Text decoding (1-5ms), result aggregation (5-20ms)

Network Backbone: InfiniBand (1-2μs) vs Ethernet (μs-ms range)

Configuration	Primary Overhead	Latency Range	Hardware Cause
Single Node	Memory bandwidth	20-80ms/layer	HBM weight loading
Data Parallel	Node synchronization	50-200ms	Coordination barriers
Model Parallel	Network communication	5-500ms/layer	Inter-node data transfer

Optimization Techniques

❑ **Quantization - Reducing Numerical Precision: (**Typically <2% performance degradation**)**

FP32 → INT8: 4x memory reduction, 2-3x speed improvement

FP32 → INT4: 8x memory reduction, significant speed gains

❑ **Pruning - Removing Parameters: (**50-90% parameter reduction maintaining performance**)**

Unstructured: Remove individual weights based on magnitude

Structured: Remove entire neurons, attention heads, or layers

❑ **Distillation - Knowledge Transfer: (**Large "teacher" model trains smaller "student" model**)**

Example: DistilBERT (66M params) achieves 97% of BERT (110M params) performance

Trade-offs: All methods balance model size, inference speed, and output quality

LAUNCHING REGIONAL MODELS

Market Opportunity Analysis

Market Size Context

- **Pakistani digital economy**: PKR 2.8 trillion with 34% annual growth
- **Current language gap**: Regional language AI market largely unaddressed
- **Long-term opportunity**: Estimated PKR 580 billion market by 2028

Potential Revenue Models

- **Enterprise Solutions**
- **Government Contracts**
- **API Services**

Potential Market Applications

Education Sector

Target market: 35M Urdu-medium students currently underserved

- ✓ Curriculum-aligned content generation
- ✓ Personalized learning explanations
- ✓ Support for students struggling with English-medium materials

Agriculture

Target users: 25M farming families needing localized information

- ✓ **Information gaps**: Weather, crop, and market data rarely available in local languages
- ✓ **Economic potential**: Optimized farming advice could reduce input costs

Conclusion

Local LLMs are the way forward!

Thank You!