Excitements in Natural Language Processing

Parameswari Krishnamurthy

param.krishna@iiit.ac.in
IIIT-Hyderabad

SRM AMARAVATI
ACM ROCS
February 21st, 2026

Natural Language

(The Human Reality)

- Unstructured & Ambiguous
- Cultural, Contextual, & Emotional
- Constantly Evolving (Code-mixing, slang)

Natural Language... and Why "Processing"?

Natural Language
(The Human Reality)

- Unstructured & Ambiguous
- Cultural, Contextual, & Emotional
- Constantly Evolving (Code-mixing, slang)

Processing
(The Scientific Challenge)

- Translating human chaos into mathematical order (Vectors, Trees, Probabilities)
- Operating at Scale (Millions of documents in seconds)
- Enabling Interaction (Making machines talk back)

The Gap

Natural Language... and Why "Processing"?

Natural Language
(The Human Reality)

- Unstructured & Ambiguous
- Cultural, Contextual, & Emotional
- Constantly Evolving (Code-mixing, slang)

Processing
(The Scientific Challenge)

- Translating human chaos into mathematical order (Vectors, Trees, Probabilities)
- Operating at Scale (Millions of documents in seconds)
- Enabling Interaction (Making machines talk back)

The Gap

Computers only "think" in math and binary.

History of Language Technology

- Invent of Scripts

History of Language Technology

- Invent of Scripts
- Writing and Publishing (information record and dissemination)

History of Language Technology

- Invent of Scripts
- Writing and Publishing (information record and dissemination)
- Broadcasting (radio, television, etc.)

History of Language Technology

- Invent of Scripts
- Writing and Publishing (information record and dissemination)
- Broadcasting (radio, television, etc.)
- Internet technology

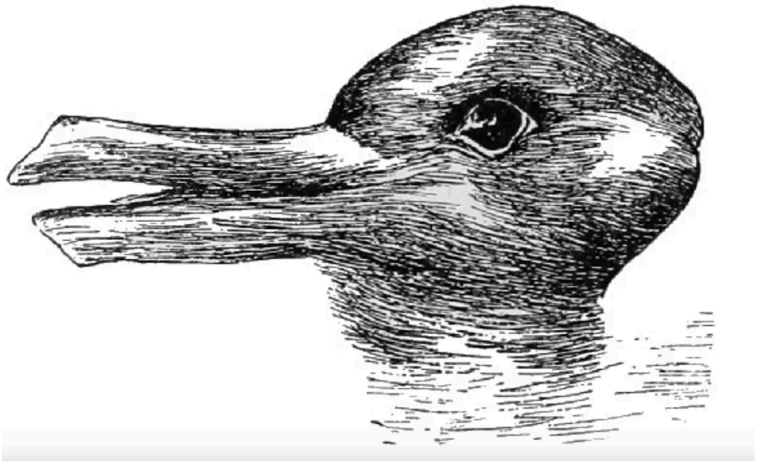
History of Language Technology

- Invent of Scripts
- Writing and Publishing (information record and dissemination)
- Broadcasting (radio, television, etc.)
- Internet technology
- Natural Language Processing (NLP) & Machine translation

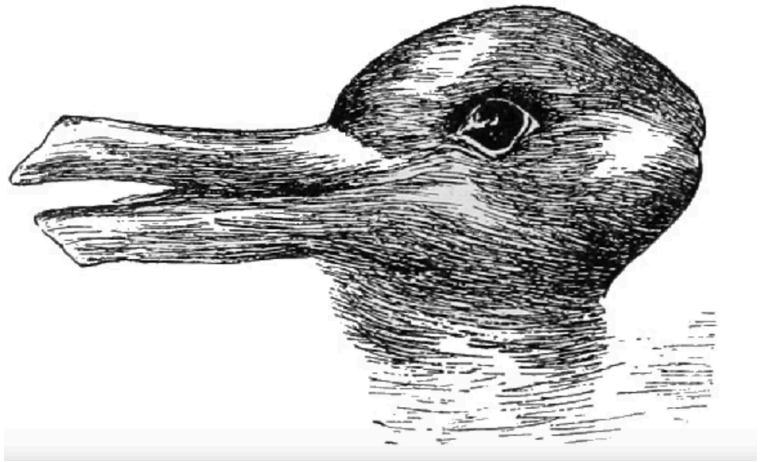
History of Language Technology

- Invent of Scripts
- Writing and Publishing (information record and dissemination)
- Broadcasting (radio, television, etc.)
- Internet technology
- Natural Language Processing (NLP) & Machine translation
- (Large) Language Models (Multimodel Multilingual NLP)

Lexical Ambiguity: What do you see?

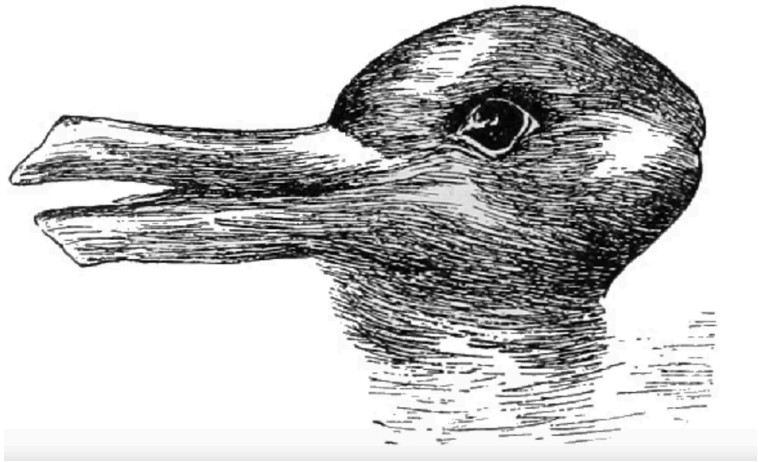


Lexical Ambiguity: What do you see?



The Duck or the Rabbit?

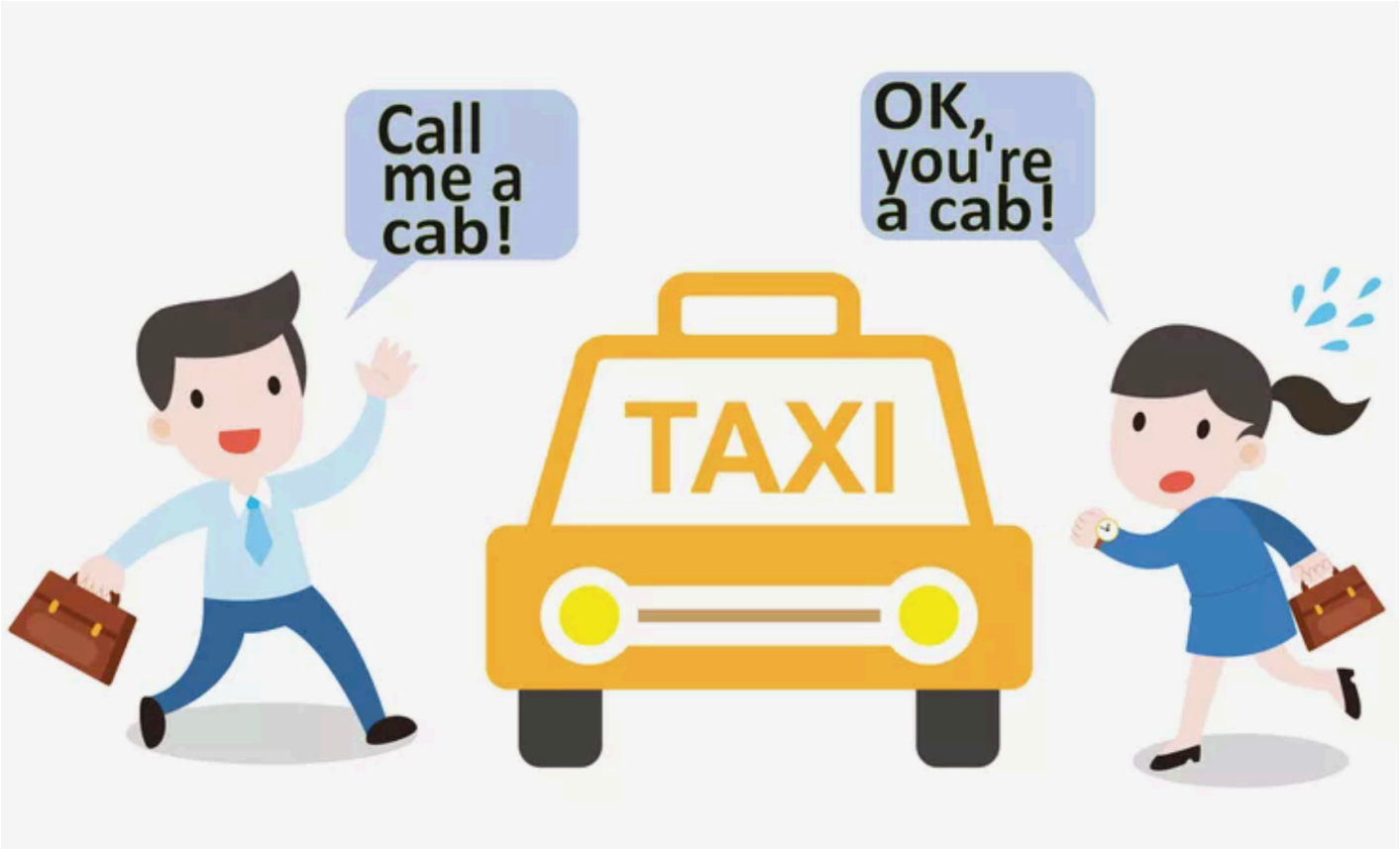
Lexical Ambiguity: What do you see?



The Duck or the Rabbit?

- **Human Perception:** We can easily flip between two distinct interpretations of the exact same visual input.

Ambiguity: When Grammar Misleads



- **The NLP Equivalent (Polysemy):** Words that look and sound identical but carry completely different meanings.
  - *English:* "Bank" (River vs. Financial Institution)

- **The NLP Equivalent (Polysemy):** Words that look and sound identical but carry completely different meanings.
  - *English:* "Bank" (River vs. Financial Institution)
  - *Hindi:* *mere paas sona hai* 'I have gold' vs. *mujhe sona hai* 'I want to sleep

- **The NLP Equivalent (Polysemy):** Words that look and sound identical but carry completely different meanings.
  - *English:* "Bank" (River vs. Financial Institution)
  - *Hindi:* *mere paas sona hai* 'I have gold' vs. *mujhe sona hai* 'I want to sleep
- **The Solution:** NLP models require **surrounding context** to lock into the correct meaning.

Syntactic Ambiguity: Attachment

కొత్త	బడి	పిల్లలు
kotha	baDi	pillalu
New	*School*	*Children*

Interpretation 1:
[కొత్త బడి] పిల్లలు → The children of the **new school**.

Interpretation 2:
కొత్త [బడి పిల్లలు] → The **new students** (school-children).

The Pillar of NLP: The Era of "Multi-"

The New Forces Reshaping Language AI

Multimodal

- Text
- Speech
- Vision

Modern NLP

Language AI is no longer single-task, single-language, or single-modality.

The Pillar of NLP: The Era of "Multi-"

The New Forces Reshaping Language AI

Multimodal
Text ▪ Speech
▪ Vision

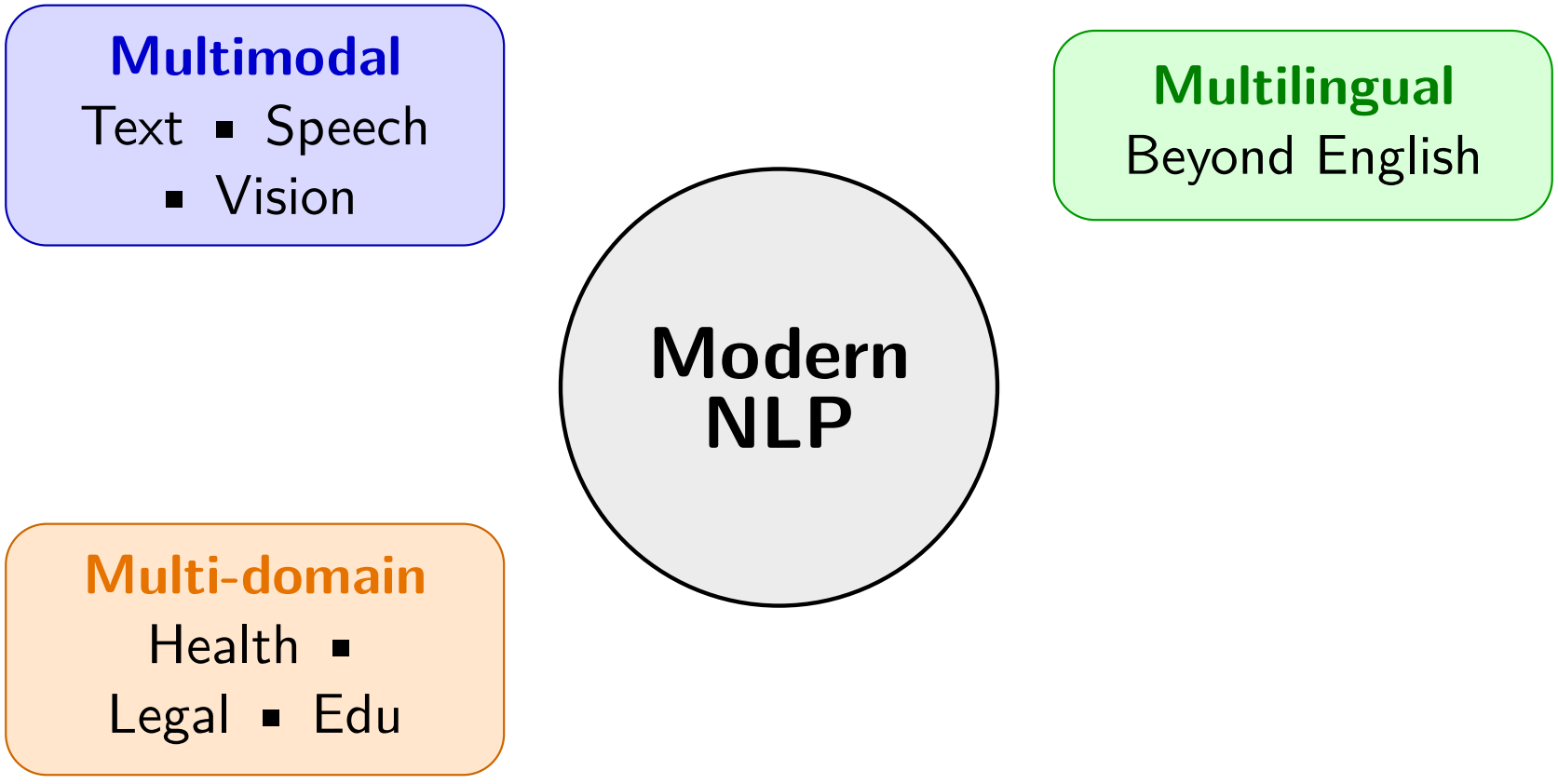
Multilingual
Beyond English

Modern NLP

Language AI is no longer single-task, single-language, or single-modality.

The Pillar of NLP: The Era of "Multi-"

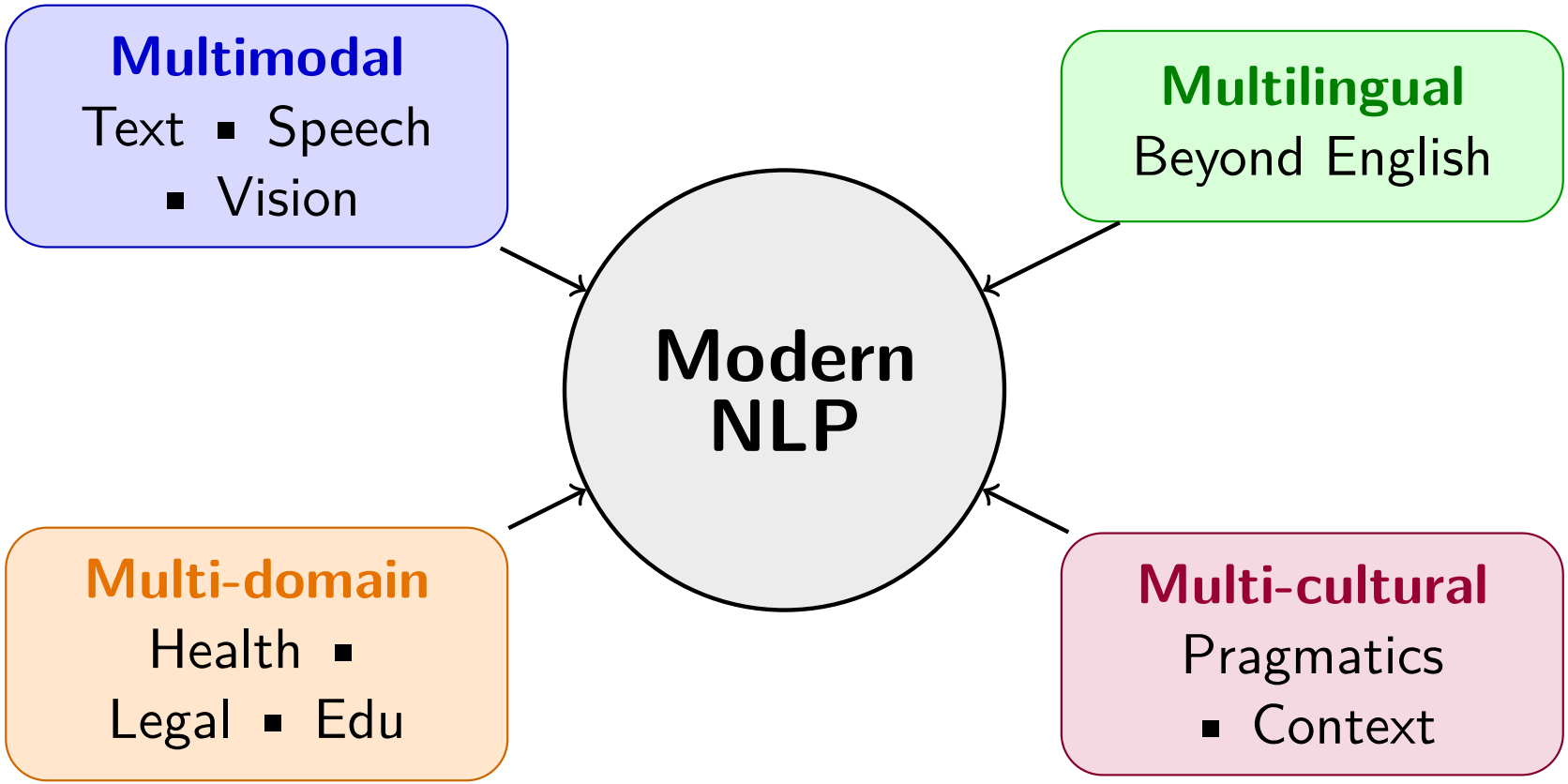
The New Forces Reshaping Language AI



Language AI is no longer single-task, single-language, or single-modality.

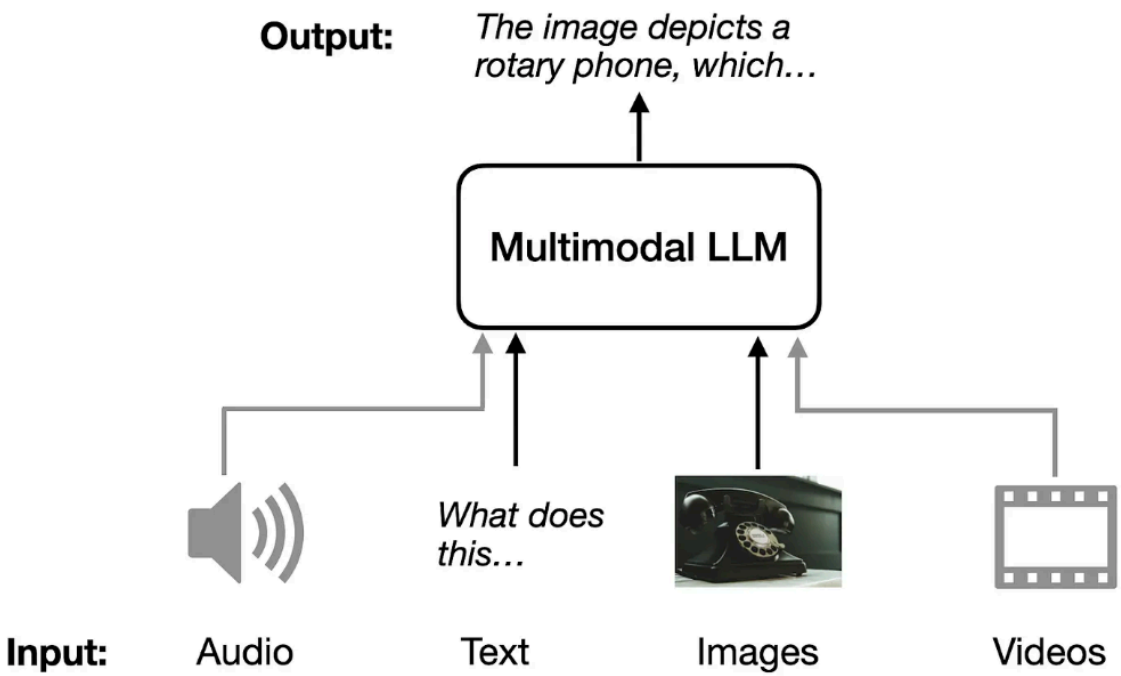
The Pillar of NLP: The Era of "Multi-"

The New Forces Reshaping Language AI



Language AI is no longer single-task, single-language, or single-modality.

1. Multimodal



Text ▪ Speech ▪ Vision

Why Text-Only NLP Fails: The "Ross Geller" Problem

Why Text-Only NLP Fails: The "Ross Geller" Problem

"I'm fine!"

1. Text (Lexical Layer)

Transcript: "I am fine."
Sentiment Positive (100%)
Result: Completely wrong.

Why Text-Only NLP Fails: The "Ross Geller" Problem

"I'm fine!"

1. Text (Lexical Layer)

Transcript: "I am fine."
Sentiment Positive (100%)
Result: Completely wrong.

2. Audio (Acoustic Layer)

Features: Abnormally high pitch, voice cracking
Audio NLP Prediction: High Stress / Panic

Why Text-Only NLP Fails: The "Ross Geller" Problem

"I'm fine!"

1. Text (Lexical Layer)

Transcript: "I am fine."
Sentiment Positive (100%)
Result: Completely wrong.

2. Audio (Acoustic Layer)

Features: Abnormally high pitch, voice cracking
Audio NLP Prediction: High Stress / Panic

3. Video (Visual Layer)

Features: Wide/manic eyes, frantic gestures, forced smile.
Vision Prediction: Distress / Unstable

The Multimodal Truth

Human communication relies heavily on **contradiction**.

The Multimodal Truth

Human communication relies heavily on **contradiction**.

When the audio and video modalities directly contradict the text modality, it usually indicates sarcasm, irony or panic.

Mapping the Modalities: Where does meaning live?

1. Text (Lexical)

What it captures:
The literal, structural foundation.

- Vocabulary & Syntax
- Core Semantics
- Named Entities
- *Constraint:* Highly literal; lacks nuance.

Mapping the Modalities: Where does meaning live?

1. Text (Lexical)

What it captures:
The literal, structural foundation.

- Vocabulary & Syntax
- Core Semantics
- Named Entities
- *Constraint:* Highly literal; lacks nuance.

2. Audio (Acoustic)

What it captures:
The delivery and state of mind.

- **Prosody:** Pitch, intonation, speech rate.
- **Disfluency:** Pauses, stutters, filler words (*um, ah*).
- Vocal intensity (shouting vs. whispering).

Mapping the Modalities: Where does meaning live?

1. Text (Lexical)

What it captures:
The literal, structural foundation.

- Vocabulary & Syntax
- Core Semantics
- Named Entities
- *Constraint:* Highly literal; lacks nuance.

2. Audio (Acoustic)

What it captures:
The delivery and state of mind.

- **Prosody:** Pitch, intonation, speech rate.
- **Disfluency:** Pauses, stutters, filler words (*um, ah*).
- Vocal intensity (shouting vs. whispering).

3. Video (Visual)

What it captures:
The physical context and grounding.

- **Gestures:** Pointing, head nods.
- Facial expressions.
- Eye contact / Gaze.
- Lip movements (helps clarify noisy audio).

The Pragmatic Layer (Where NLP struggles)

Complex human communication requires aligning all three modalities:

- **Emotion:** Can be stated in text, but is heavily validated by audio (tone) and video (tears/smiles).
- **Sarcasm & Irony:** Often occurs when modalities contradict (e.g., the *Text* says "Great job", but the *Audio* is flat and the *Video* shows an eye-roll).

2. Multilingual

22 Scheduled Languages ▪ 100+ Languages ▪ 1000+ Varieties

The English-Centric Reality

- Most large language models are trained predominantly on English web data.

The English-Centric Reality

- Most large language models are trained predominantly on English web data.

The English-Centric Reality

- Most large language models are trained predominantly on English web data.
- High-resource languages dominate:
  - Data availability
  - Benchmarks
  - Evaluation metrics

The English-Centric Reality

- Most large language models are trained predominantly on English web data.
- High-resource languages dominate:
  - Data availability
  - Benchmarks
  - Evaluation metrics

The English-Centric Reality

- Most large language models are trained predominantly on English web data.
- High-resource languages dominate:
  - Data availability
  - Benchmarks
  - Evaluation metrics
- Low-resource languages are digitally under-represented.

Implication

Multilingual $\neq$ Equal Representation.

The “Token Tax” Problem

- Standard tokenizers are optimized for English morphology.

The “Token Tax” Problem

- Standard tokenizers are optimized for English morphology.
- Indic scripts often fragment into multiple subword tokens.

The “Token Tax” Problem

- Standard tokenizers are optimized for English morphology.
- Indic scripts often fragment into multiple subword tokens.

Example:

- English word $\rightarrow$ 1 token

The “Token Tax” Problem

- Standard tokenizers are optimized for English morphology.
- Indic scripts often fragment into multiple subword tokens.

Example:

- English word $\rightarrow$ 1 token
- Hindi$\rightarrow$ 2-3 tokens

The "Token Tax" Problem

- Standard tokenizers are optimized for English morphology.
- Indic scripts often fragment into multiple subword tokens.

Example:

- English word $\rightarrow$ 1 token
- Hindi$\rightarrow$ 2-3 tokens
- Tamil /Telugu word $\rightarrow$ 3–5 or more tokens

Consequence

Higher inference cost + Reduced effective context window.

Complex Morpheme Segmentation in Some Languages

- Some languages require complex morpheme segmentation.

Turkish:

- Uygarlastiramadiklarimizdanmissinizcasina
  '(behaving) as if you are among those whom we could not civilize'
- Uygar 'civilized' + las 'become' + tir 'cause' + ama 'not able' + dik 'past' + lar 'plural' + imiz '1pl' + dan 'ablative' + mis 'past' + siniz '2pl' + casina 'as if'

Example for derivation from Telugu:

పగలగొట్టించిపెట్టమనినివ్వదలచుకోలేకపోతున్నాను
pagalagottiMcipettamananivvadalacukooleekapootunnaanu

pagulu+a-kottu+iMcu+i-pettu+a-manu+a-ivvu+a-daluvu+i-konu+a-leeka-poo+ tunn+1,sg,any
break+inf-strike+cause+cpm-benefactive+inf-tell+inf-permit+inf-think+cpm-reflexive+inf-neg+go+prog+1, sg

'I could not think to permit someone to tell for my sake to break something'
(pc, G. Uma Maheshwar Rao)

Tokenizer Behavior on a Morphologically Rich Word

పగలగోట్టించిపెట్టమననివ్వదలచుకోలేకపోతున్నాను

Tokenizer	**Segmentation**
BPE	పగల + గో + ట్టి + ◌ించి + పెట్ట + మన + నివ్వ + దల + చుకో + లేకపో + తున్నాను

Tokenizer Behavior on a Morphologically Rich Word

పగలగోట్టించిపెట్టమననివ్వదలచుకోలేకపోతున్నాను

Tokenizer	**Segmentation**
BPE	పగల + గో + ట్టి + ించి + పెట్ట + మన + నివ్వ + దల + చుకో + లేకపో + తున్నాను
UnigramLM	పగల + గో + ట్టి + ించి + పెట్ట + మన + నివ్వ + దలచ + ుకోలేక + పోతున్నా + ను

Tokenizer Behavior on a Morphologically Rich Word

పగలగోట్టించిపెట్టమననివ్వదలచుకోలేకపోతున్నాను

Tokenizer	**Segmentation**
BPE	పగల + గో + ట్టి + ించి + పెట్ట + మన + నివ్వ + దల + చుకో + లేకపో + తున్నాను
UnigramLM	పగల + గో + ట్టి + ించి + పెట్ట + మన + నివ్వ + దలచ + ుకోలేక + పోతున్నా + ను
WordPiece	పగల + గో + ట్టి + ించి + పెట్ట + మన + నివ్వ + దల + చుకో + లేకపోతున్న + ాను

Tokenizer Behavior on a Morphologically Rich Word

పగలగోట్టించిపెట్టమననివ్వదలచుకోలేకపోతున్నాను

Tokenizer	**Segmentation**
BPE	పగల + గో + ట్టి + ించి + పెట్ట + మన + నివ్వ + దల + చుకో + లేకపో + తున్నాను
UnigramLM	పగల + గో + ట్టి + ించి + పెట్ట + మన + నివ్వ + దలచ + ుకోలేక + పోతున్నా + ను
WordPiece	పగల + గో + ట్టి + ించి + పెట్ట + మన + నివ్వ + దల + చుకో + లేకపోతున్న + ాను
SSLM	ప + గల + గొ + ట్ట + ించ + ి + పెట్ట + మ + నని + వ్వ + దల + చుక + ోలే + కపోత + ున్న + ాను

Tokenizer Behavior on a Morphologically Rich Word

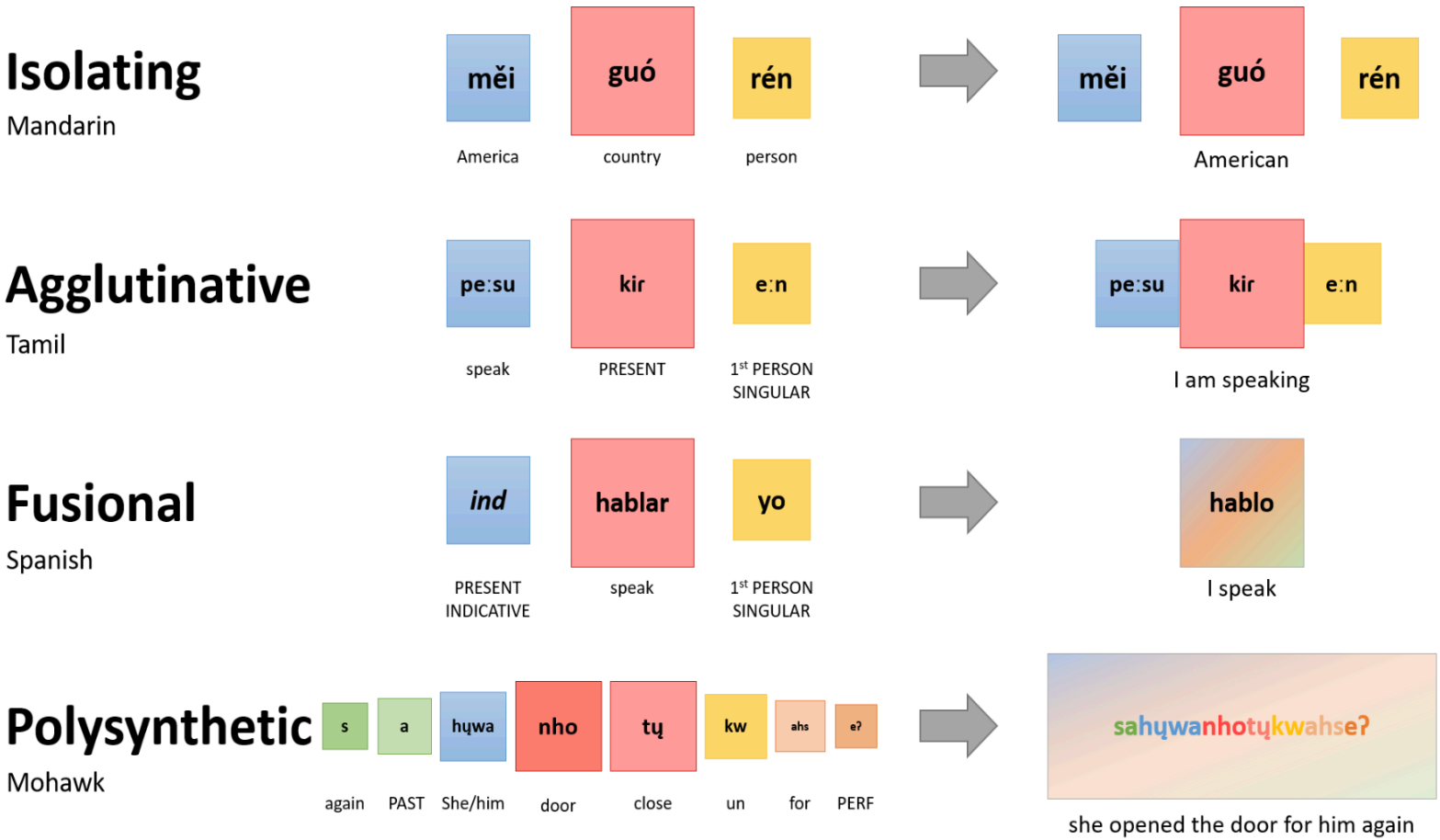
పగలగోట్టించిపెట్టమననివ్వదలచుకోలేకపోతున్నాను

Tokenizer	Segmentation
BPE	పగల + గో + ట్టి + ించి + పెట్ట + మన + నివ్వ + దల + చుకో + లేకపో + తున్నాను
UnigramLM	పగల + గో + ట్టి + ించి + పెట్ట + మన + నివ్వ + దలచ + ుకోలేక + పోతున్నా + ను
WordPiece	పగల + గో + ట్టి + ించి + పెట్ట + మన + నివ్వ + దల + చుకో + లేకపోతున్న + ాను
SSLM	ప + గల + గొ + ట్ట + ించ + ి + పెట్ట + మ + నని + వ్వ + దల + చుక + ోలే + కపోత + ున్న + ాను

Observation

Subword tokenizers fragment differently — none align perfectly with linguistic morphemes.

Morphological Typology



The "English-Centric" Reality

- **The Data Imbalance:**
  - Global LLMs are trained primarily on English data (Common Crawl $> 45\%$).

The "English-Centric" Reality

- **The Data Imbalance:**
  - Global LLMs are trained primarily on English data (Common Crawl $> 45\%$).
  - Indic languages are "low-resource" in the digital domain, despite 1.4B+ speakers.

The "English-Centric" Reality

- **The Data Imbalance:**
  - Global LLMs are trained primarily on English data (Common Crawl $> 45\%$).
  - Indic languages are "low-resource" in the digital domain, despite 1.4B+ speakers.
- **The "Token Tax":**
  - Standard tokenizers (GPT-4, Llama 2) fracture Indic scripts.

The "English-Centric" Reality

- **The Data Imbalance:**
  - Global LLMs are trained primarily on English data (Common Crawl $> 45\%$).
  - Indic languages are "low-resource" in the digital domain, despite 1.4B+ speakers.
- **The "Token Tax":**
  - Standard tokenizers (GPT-4, Llama 2) fracture Indic scripts.
  - **Example:** A Telugu word might split into 4-5 tokens, while the English equivalent is 1 token.

The "English-Centric" Reality

- **The Data Imbalance:**
  - Global LLMs are trained primarily on English data (Common Crawl $> 45\%$).
  - Indic languages are "low-resource" in the digital domain, despite 1.4B+ speakers.
- **The "Token Tax":**
  - Standard tokenizers (GPT-4, Llama 2) fracture Indic scripts.
  - **Example:** A Telugu word might split into 4-5 tokens, while the English equivalent is 1 token.
  - **Consequence:** Higher inference costs and reduced context window for Indian languages.

The "English-Centric" Reality

- **The Data Imbalance:**
  - Global LLMs are trained primarily on English data (Common Crawl $> 45\%$).
  - Indic languages are "low-resource" in the digital domain, despite 1.4B+ speakers.
- **The "Token Tax":**
  - Standard tokenizers (GPT-4, Llama 2) fracture Indic scripts.
  - **Example:** A Telugu word might split into 4-5 tokens, while the English equivalent is 1 token.
  - **Consequence:** Higher inference costs and reduced context window for Indian languages.

Linguistic Diversity as a Challenge

- **Script Complexity:**
  - Indic scripts are Abugidas (Alpha-syllabary).

Linguistic Diversity as a Challenge

- **Script Complexity:**
  - Indic scripts are Abugidas (Alpha-syllabary).
  - Complex conjuncts (e.g., *ksha*, *tra*) pose challenges for standard Byte-Pair Encoding (BPE).

Linguistic Diversity as a Challenge

- **Script Complexity:**
  - Indic scripts are Abugidas (Alpha-syllabary).
  - Complex conjuncts (e.g., *ksha*, *tra*) pose challenges for standard Byte-Pair Encoding (BPE).
- **Code-Mixing:**
  - We rarely speak "pure" language.

Linguistic Diversity as a Challenge

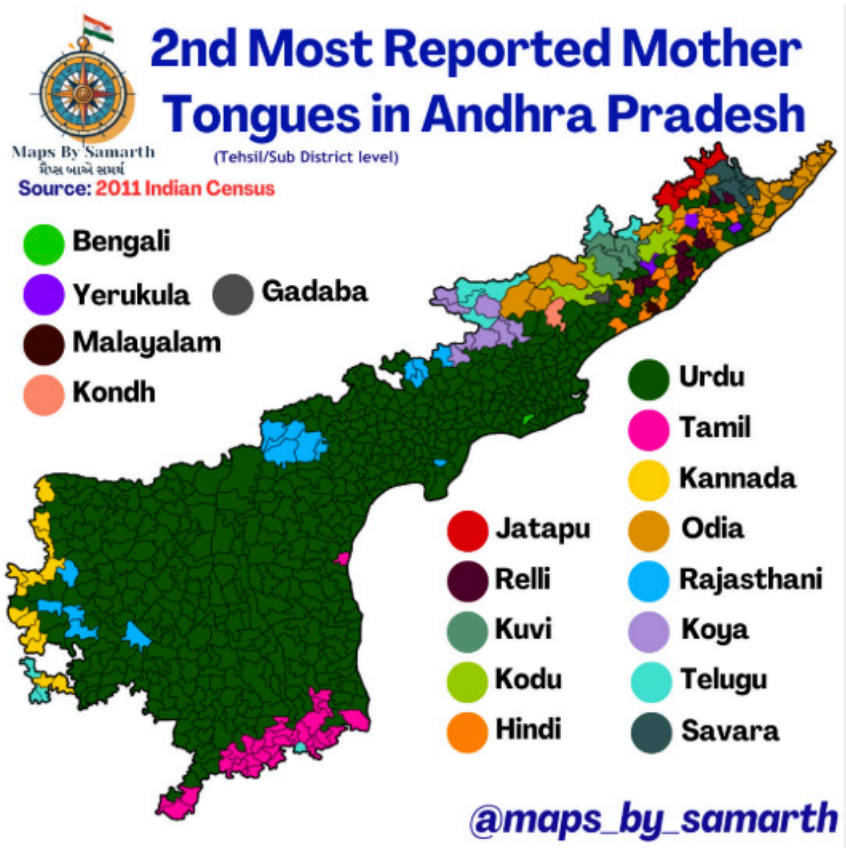
- **Script Complexity:**
  - Indic scripts are Abugidas (Alpha-syllabary).
  - Complex conjuncts (e.g., *ksha*, *tra*) pose challenges for standard Byte-Pair Encoding (BPE).
- **Code-Mixing:**
  - We rarely speak “pure” language.
  - ”Hinglish” (Hindi+English) or ”Tanglish” (Tamil+English) is the norm in digital communication.

Linguistic Diversity as a Challenge

- **Script Complexity:**
  - Indic scripts are Abugidas (Alpha-syllabary).
  - Complex conjuncts (e.g., *ksha*, *tra*) pose challenges for standard Byte-Pair Encoding (BPE).
- **Code-Mixing:**
  - We rarely speak “pure” language.
  - ”Hinglish” (Hindi+English) or ”Tanglish” (Tamil+English) is the norm in digital communication.

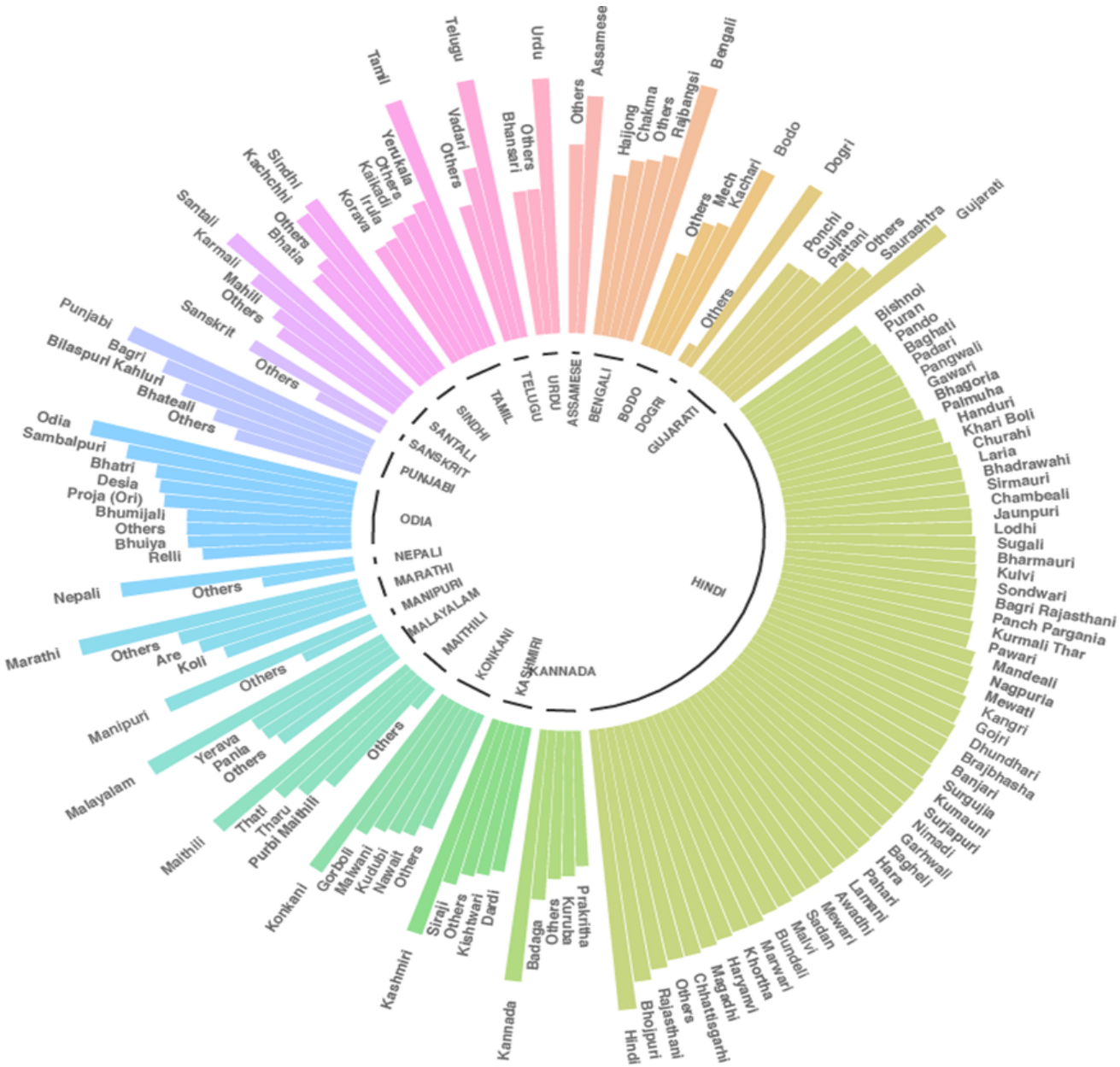
Mulitilingualism in use

Tiruchirappalli International Airport(Tamil-Hindi-English)

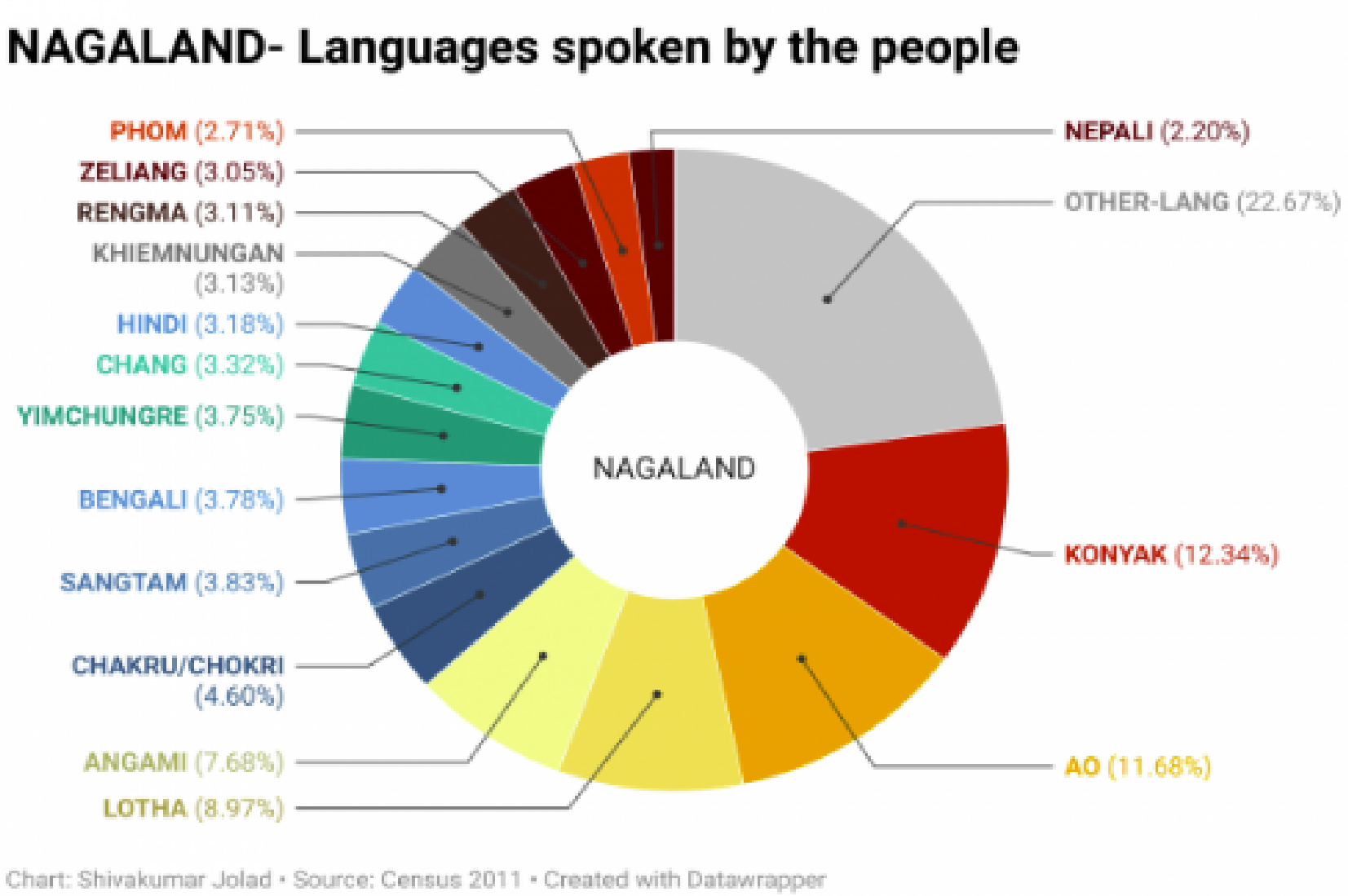


Language Map of AP

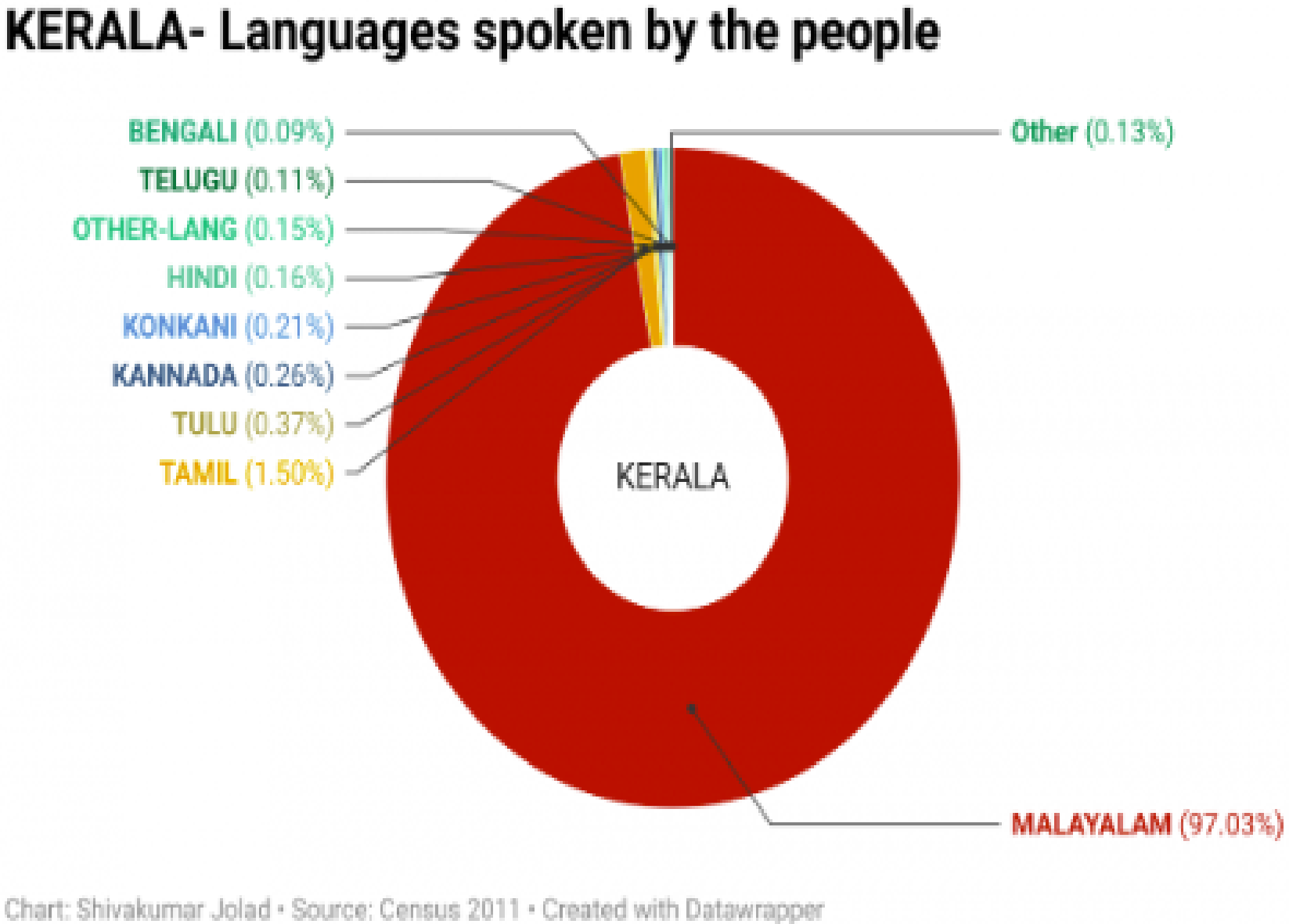
Indian Languages



Most Diverse Indian State



Least Diverse Indian State



Languages of India

Indo-Aryan Family: Assamese , Baigani, Banjari, Bengali „ Bhatri, Bhili, Bhunjia, Chakma, Chhattisgarhi, Dhanki, Dhodia, Dhundhari, Gadiali, Gamit/Gavti, Garasia/Girasia, Gojri/Gujjari, Gujarati , Hajong, Halbi, Harauti, Hindi , Jaunsari, Kachchi, Khotta, Kinnui, Kokni, Konkani , Kotwalia, Kudamamali, Thar, Lambani or Lamani , Laria, Magahi, Mahl, Marathi , Mavchi, Mewnri, Nagpuri, Naikadi, Nimari, Oriya , Rathi, Sarhodi, Shina, Tharu, Wagri, Warli.

Languages of India

Tibeto-Burman Family: Adi Ashing, Adi Bokar, Adi Bori, Adi Gallong, Adi Komkar, Adi Milang, Adi Minyong, Adi Padam, Adi Karko, Pailibo, Adi Pangi, Adi Pasi, Adi Ramo, Adi Shimong, Adi Tangam, Aimol, Anal, Angami, Ao, Apatani, Balti, Bangni/Dafla, Bawm, Bhotia, Biate, Bodo, Bugun, Chakhesang, Champa, Chang, Chiru, Chote, Chung, Dalu, Deori, Dokpa/Droskat, Duhlian-Twang, Gangte, Garo, Halam, Hmar, Hrusso/Aka, Hualngo, Kabui, Kachari, Kagati, Kak barak, Khamba, Khampa, Khiamngan, Koch, Koireng, Konyak, Kuki, Ladakhi, Lahauli, Lai Hawlh, Lakher/Mara, Lalung, Lamgang, Lepcha, Lisu, Lotha, Lushai/Mizo, Mag/Mogh, Mao, Maram, Maring, Memba, Mikir, Miri, Mishing, Mishmi, Monpa, Monsang, Moyon, Na, Naga, Sherdukpen, Nishi, Nocte, Paite, Pang, Phom, Pochury, Ralte, Rengma, Riang, Sajalong/Miju, Sangtam, Sema, Sherpa, Singpho, Sulung, Tagin, Tangsa, Thado, Tangkhul, Tibetan, Toto, Vaiphei, Wancho, Yim-chungre, Zakhring/Meyer, Zemi, Zou.

Languages of India

Dravidian Family: Dhurwa, Gadaba tribe , Gondi, Kadar tribe, Kannada, Kodagu, Kolami, Koraga, Kota, Koya/Koi, Kui, Kurukh, Kuvi, Malayalam, Malta, Maria, Naiki, Parji, Pengo, Tamil, Telugu , Toda, Tulu, Yerukula.
Austro-Asiatic Family: Asuri, Bhumij tribe, Birhor tribe, Birjia tribe, Bondo, Diday, Gutob, Ho, Juang, Kharia, Khasi, Kherwari, Korku, Korwa, Kurmi, Lodha, Mundari, Nicobarese, Santali, Saora/Savara, Shompen, Thar.
Andamanese Family: Andamanese Tribe, Jarawa tribe, Onge, Santinelese.

Indian Language and Data Resources

Language	Code	Pop. (M)	CC Size (%)	Cat.
English	en	1,452	45.8786	H
Russian	ru	258	5.9692	H
German	de	134	5.8811	H
Chinese	zh	1,118	4.8747	H
Japanese	jp	125	4.7884	H
French	fr	274	4.7254	H
Spanish	es	548	4.4690	H
Italian	it	68	2.5712	H
Dutch	nl	30	2.0585	H
Polish	pl	45	1.6636	H
Portuguese	pt	257	1.1505	H
Vietnamese	vi	85	1.0299	H
Turkish	tr	88	0.8439	M
Indonesian	id	199	0.7991	M
Swedish	sv	13	0.6969	M
Arabic	ar	274	0.6658	M
Persian	fa	130	0.6582	M
Korean	ko	81	0.6498	M
Greek	el	13	0.5870	M
Thai	th	60	0.4143	M
Ukrainian	uk	33	0.3304	M
Bulgarian	bg	8	0.2900	M
Hindi	hi	602	0.1588	M

Indian Language and Data Resources

Bengali	bn	272	0.0930	L
Tamil	ta	86	0.0446	L
Urdu	ur	231	0.0274	L
Malayalam	ml	36	0.0222	L
Marathi	mr	99	0.0213	L
Telugu	te	95	0.0183	L
Gujarati	gu	62	0.0126	L
Burmese	my	33	0.0126	L
Kannada	kn	64	0.0122	L
Swahili	sw	71	0.0077	X
Punjabi	pa	113	0.0061	X
Kyrgyz	ky	5	0.0049	X
Odia	or	39	0.0044	X
Assamesese	as	15	0.0025	X

Table 1: List of languages, language codes, numbers of first and second speakers, data ratios in the CommonCrawl corpus, and language categories. The languages are grouped into categories based on their data ratios in the CommomCrawl corpus: High Resource (H, $> 1\%$), Medium Resource (M, $> 0.1\%$), and Low Resource (L, $> 0.01\%$), and Extremely-Low Resource (X, $< 0.01\%$).

Key Takeaways

- Language is **not just data** — it is structure, culture, and use.

Key Takeaways

- Language is **not just data** — it is structure, culture, and use.
- Indian languages pose challenges that go beyond scale:
  - Rich morphology and agglutination

Key Takeaways

- Language is **not just data** — it is structure, culture, and use.
- Indian languages pose challenges that go beyond scale:
  - Rich morphology and agglutination
  - Script and orthographic complexity

Key Takeaways

- Language is **not just data** — it is structure, culture, and use.
- Indian languages pose challenges that go beyond scale:
  - Rich morphology and agglutination
  - Script and orthographic complexity
  - Code-mixing and spoken-first usage

Key Takeaways

- Language is **not just data** — it is structure, culture, and use.
- Indian languages pose challenges that go beyond scale:
  - Rich morphology and agglutination
  - Script and orthographic complexity
  - Code-mixing and spoken-first usage
- LLMs inherit linguistic assumptions from their training data.

Key Takeaways

- Language is **not just data** — it is structure, culture, and use.
- Indian languages pose challenges that go beyond scale:
  - Rich morphology and agglutination
  - Script and orthographic complexity
  - Code-mixing and spoken-first usage
- LLMs inherit linguistic assumptions from their training data.

What Has Changed Recently?

- Shift from **adaptation** to **native training**

What Has Changed Recently?

- Shift from **adaptation** to **native training**
- Emergence of Indic-first models:
  - IndicTrans2, BhashaVerse

What Has Changed Recently?

- Shift from **adaptation** to **native training**
- Emergence of Indic-first models:
  - IndicTrans2, BhashaVerse
  - Airavata, OpenHathi, Sarvam, BharatGen etc.

What Has Changed Recently?

- Shift from **adaptation** to **native training**
- Emergence of Indic-first models:
  - IndicTrans2, BhashaVerse
  - Airavata, OpenHathi, Sarvam, BharatGen etc.
- Better tokenization strategies:
  - Script-aware tokenizers
  - Improved fertility scores (token $\approx$ word)

What Has Changed Recently?

- Shift from **adaptation** to **native training**
- Emergence of Indic-first models:
  - IndicTrans2, BhashaVerse
  - Airavata, OpenHathi, Sarvam, BharatGen etc.
- Better tokenization strategies:
  - Script-aware tokenizers
  - Improved fertility scores (token $\approx$ word)

3. Multi-domain

Beyond General Chat ▪ Towards Specialized Intelligence

Why General-Purpose LLMs Are Not Enough

- Trained on broad web-scale data

Why General-Purpose LLMs Are Not Enough

- Trained on broad web-scale data
- Optimized for fluency and plausibility

Why General-Purpose LLMs Are Not Enough

- Trained on broad web-scale data
- Optimized for fluency and plausibility
- Not optimized for domain-specific precision

Core Problem

Fluent text $\neq$ Correct domain knowledge.

Domains Where Errors Have Consequences

Healthcare

- **Negation Error:** "No evidence of malignancy" $\rightarrow$ "Evidence of malignancy"
- **Dosage Misinterpretation:** "Take 1 tablet twice daily" $\rightarrow$ "Take 2 tablets daily"
- **Clinical Summary Error:** Dropping allergy information in discharge note

Domains Where Errors Have Consequences

Healthcare

- **Negation Error:** "No evidence of malignancy" → "Evidence of malignancy"
- **Dosage Misinterpretation:** "Take 1 tablet twice daily" → "Take 2 tablets daily"
- **Clinical Summary Error:** Dropping allergy information in discharge note

Legal

- **Terminology Precision:** "Shall" vs "May"
- **Contract Ambiguity:** "The contractor shall deliver materials and equipment insured." (Who is insured?)
- **Policy Translation Risk:** Welfare eligibility condition mistranslated

Education / Governance / Agriculture: Incorrect crop advisory or exam instruction can mislead thousands.

LLMs for Indian Language Applications

- **Healthcare:**
  - Doctor–patient communication
  - Speech summarization and translation

LLMs for Indian Language Applications

- **Healthcare:**
  - Doctor–patient communication
  - Speech summarization and translation
- **Governance:**
  - Multilingual access to schemes and services

LLMs for Indian Language Applications

- **Healthcare:**
  - Doctor–patient communication
  - Speech summarization and translation
- **Governance:**
  - Multilingual access to schemes and services
- **Education:**
  - Mother-tongue instruction
  - Speech-based tutoring systems

LLMs for Indian Language Applications

- **Healthcare:**
  - Doctor–patient communication
  - Speech summarization and translation
- **Governance:**
  - Multilingual access to schemes and services
- **Education:**
  - Mother-tongue instruction
  - Speech-based tutoring systems
- **Assistive Technologies:**
  - Voice interfaces for low-literate users

LLMs for Indian Language Applications

- **Healthcare:**
  - Doctor–patient communication
  - Speech summarization and translation
- **Governance:**
  - Multilingual access to schemes and services
- **Education:**
  - Mother-tongue instruction
  - Speech-based tutoring systems
- **Assistive Technologies:**
  - Voice interfaces for low-literate users

Approaches to Multi-domain Modeling

- Domain-adaptive pretraining

Approaches to Multi-domain Modeling

- Domain-adaptive pretraining
- Instruction tuning with domain data

Approaches to Multi-domain Modeling

- Domain-adaptive pretraining
- Instruction tuning with domain data
- Retrieval-Augmented Generation (RAG)

Approaches to Multi-domain Modeling

- Domain-adaptive pretraining
- Instruction tuning with domain data
- Retrieval-Augmented Generation (RAG)
- Mixture-of-Experts (MoE) architectures

Trend

Shift from single monolithic models to modular expertise.

Multi-tasking Within a Domain

A single domain system may need to:

- Translate

Multi-tasking Within a Domain

A single domain system may need to:

- Translate
- Summarize

Multi-tasking Within a Domain

A single domain system may need to:

- Translate
- Summarize
- Extract entities

Multi-tasking Within a Domain

A single domain system may need to:

- Translate
- Summarize
- Extract entities
- Answer questions

Multi-tasking Within a Domain

A single domain system may need to:

- Translate
- Summarize
- Extract entities
- Answer questions
- Detect risk signals

Challenge

Can one model reliably perform all tasks?

4. Multi-cultural

Beyond Syntax ▪ Beyond Semantics ▪ Towards Pragmatics

Language Encodes Culture

- Politeness levels

Language Encodes Culture

- Politeness levels
- Honorific systems

Language Encodes Culture

- Politeness levels
- Honorific systems
- Indirectness vs directness

Language Encodes Culture

- Politeness levels
- Honorific systems
- Indirectness vs directness
- Social hierarchy markers

Language Encodes Culture

- Politeness levels
- Honorific systems
- Indirectness vs directness
- Social hierarchy markers
- Context-dependent meaning

Key Insight

Two sentences can be semantically identical but culturally inappropriate.

Honorifics and Politeness Systems

- Hindi: *tu* / *tum* / *aap*

Honorifics and Politeness Systems

- Hindi: *tu* / *tum* / *aap*
- Tamil: informal vs respectful verb morphology

Honorifics and Politeness Systems

- Hindi: *tu* / *tum* / *aap*
- Tamil: informal vs respectful verb morphology
- Telugu: verb agreement encodes respect

Honorifics and Politeness Systems

- Hindi: *tu* / *tum* / *aap*
- Tamil: informal vs respectful verb morphology
- Telugu: verb agreement encodes respect
- Japanese, Korean: hierarchical grammar

Modeling Challenge

LLMs often default to neutral tone — neutral can be disrespectful.

Indirectness and Pragmatic Meaning

- "You ate?" (Greeting, not question)

Indirectness and Pragmatic Meaning

- "You ate?" (Greeting, not question)
- "It is a bit warm here." (Request to turn on AC)

Indirectness and Pragmatic Meaning

- "You ate?" (Greeting, not question)
- "It is a bit warm here." (Request to turn on AC)
- Silence can signal disagreement

Indirectness and Pragmatic Meaning

- "You ate?" (Greeting, not question)
- "It is a bit warm here." (Request to turn on AC)
- Silence can signal disagreement

Pragmatics

Meaning is inferred, not stated.

The New NLP is Multi-Dimensional

Modern NLP must be:

- Multimodal
- Multilingual
- Multi-domain
- Multi-cultural

Final Thought

If we ignore even one dimension, we build systems that work — but only for a few.

What Must Change Going Forward

- **1. Native Data Creation**
  - Spoken, conversational, domain-specific corpora
  - Code-mixed and dialectal data

What Must Change Going Forward

- **1. Native Data Creation**
  - Spoken, conversational, domain-specific corpora
  - Code-mixed and dialectal data
- **2. Linguistically Informed Modeling**
  - Morphology-aware tokenization
  - Discourse- and pragmatics-aware systems

What Must Change Going Forward

- **1. Native Data Creation**
  - Spoken, conversational, domain-specific corpora
  - Code-mixed and dialectal data
- **2. Linguistically Informed Modeling**
  - Morphology-aware tokenization
  - Discourse- and pragmatics-aware systems
- **3. Rethinking Evaluation**
  - Beyond BLEU / ROUGE
  - Cultural adequacy and domain reliability
  - Human-in-the-loop validation

What Must Change Going Forward

- **1. Native Data Creation**
  - Spoken, conversational, domain-specific corpora
  - Code-mixed and dialectal data
- **2. Linguistically Informed Modeling**
  - Morphology-aware tokenization
  - Discourse- and pragmatics-aware systems
- **3. Rethinking Evaluation**
  - Beyond BLEU / ROUGE
  - Cultural adequacy and domain reliability
  - Human-in-the-loop validation
- **4. Responsible & Inclusive AI**
  - Language equity as a design principle
  - Bias-aware multilingual deployment

Final Message

LLMs are not multilingual, multimodal, multi-domain, or multi-cultural by default.

Final Message

LLMs are not multilingual, multimodal,
multi-domain, or multi-cultural by default.

They become inclusive only when we respect

linguistic structure, speech reality, domain knowledge, and cultural context.

Final Message

LLMs are not multilingual, multimodal,
multi-domain, or multi-cultural by default.

They become inclusive only when we respect

linguistic structure, speech reality, domain knowledge, and cultural context.

For India, this means:

Native data. Native models. Native evaluation.

Final Message

LLMs are not multilingual, multimodal,
multi-domain, or multi-cultural by default.

They become inclusive only when we respect

linguistic structure, speech reality, domain knowledge, and cultural context.

For India, this means:

Native data. Native models. Native evaluation.

Language is not just input to AI.

It is digital infrastructure.