

How to build Tamil LLM?

As a community

கணித்தமிழ்24 மாநாடு

**Abinaya Mahendiran
Feb 9, 2024**



Abinaya Mahendiran

- CTO, Nunnari Labs,
- Program Manager, IITM
- M.Tech IT, International Institute of Information and Technology Bangalore
- Volunteer at AI Tamil Nadu, WTM, Data Conversations, GHCI, WAI, Women Who Code
- Interests: Building NLP/NLU/NLG/MLOps/Gen AI systems, Open source, Applied Research

 <https://abinayam02.github.io>

 <https://www.linkedin.com/in/abinayamahendiran/>

 @freakynut

 <https://medium.com/@abinayamahendiran>

 https://topmate.io/abinaya_mahendiran

Agenda

Types of AI

Why Tamil AI?

Natural Language Processing

AI Tools using NLP

NLP: Techniques

NLP: Pipeline

Data: NLP Pipeline

Building Tamil AI - Necessities

Data: Curation Challenges

Data: Curation Framework

Core Components

Models: Traditional Approach

Models: Transfer Learning

Foundation Models (LLM)

LLM Training

Prompt Engineering

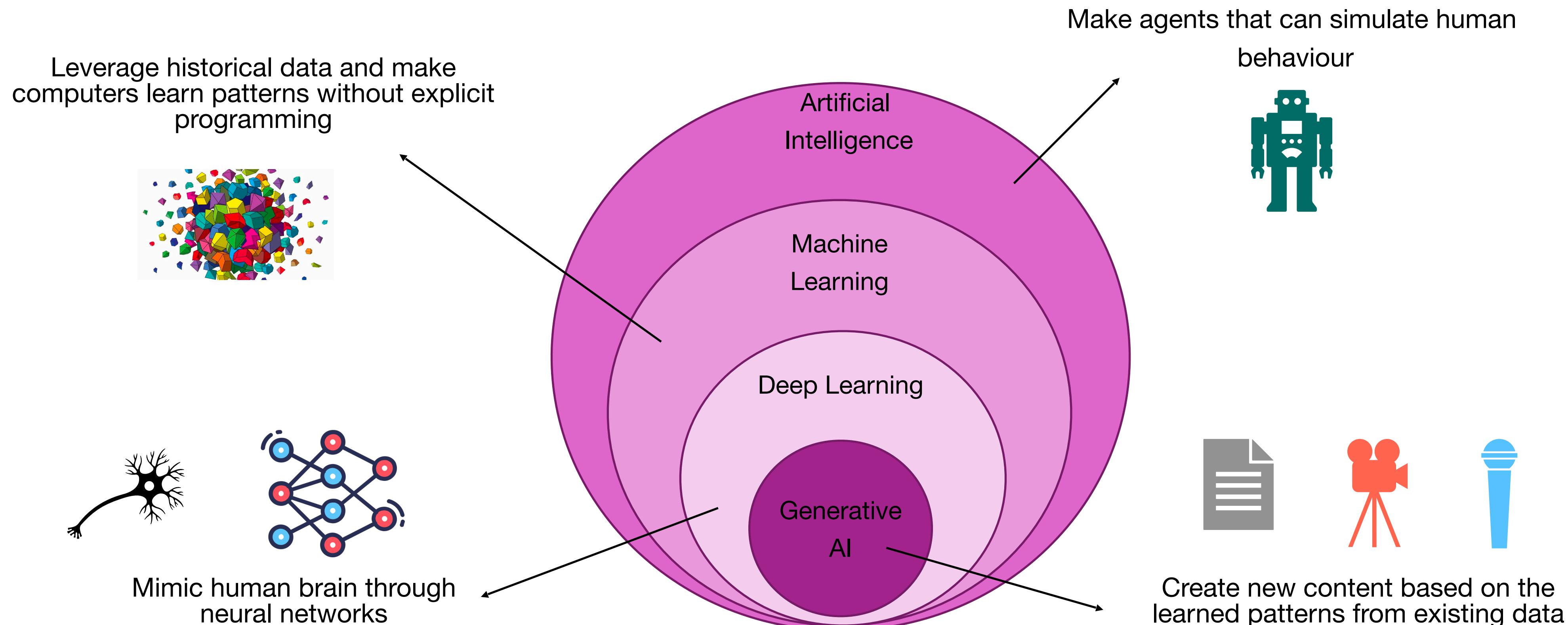
Models for Tamil

Limitations of Gen AI

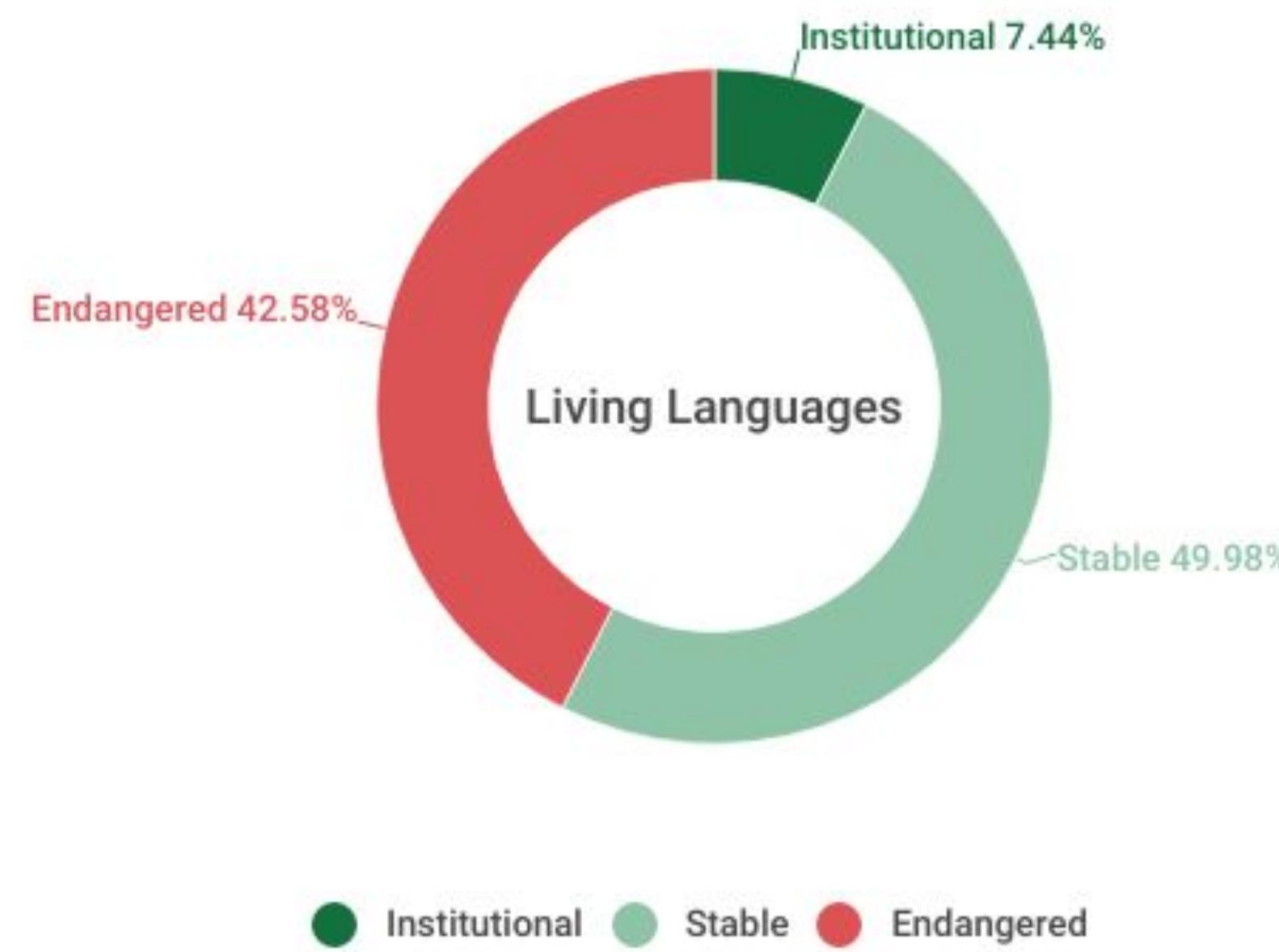
Gen AI: Research Directions

Role of the community

Types of AI



Why Tamil AI?



 Ethnologue

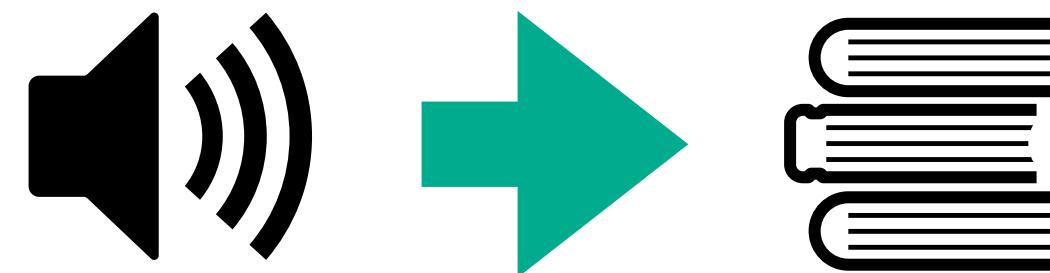
சங்கம்	கால இடைவெளி	கவிஞர்களின் எண்ணிக்கை	இராச்சியம் [9]	புத்தகங்கள் [9]
முதலில் ஆண்டுகள் [9]	4440	549 [9]	பாண்டியா	புத்தகங்கள் எதுவும் பிழைக்கவில்லை
இரண்டாவது ஆண்டுகள் [9]	3700	1700 [9]	பாண்டியா	தொல்காப்பியம் (ஆசிரியர் – தொல்காப்பியர்)
மூன்றாவது ஆண்டுகள் [9]	1850		பாண்டியா	சங்க இலக்கியம் முழுவதையும் உள்ளடக்கியது

- ~80 million people speak Tamil
- Rich literature and yet underrepresented in NLP.
- AI tools can help sustain the language and its culture.
- Future generations can benefit greatly.

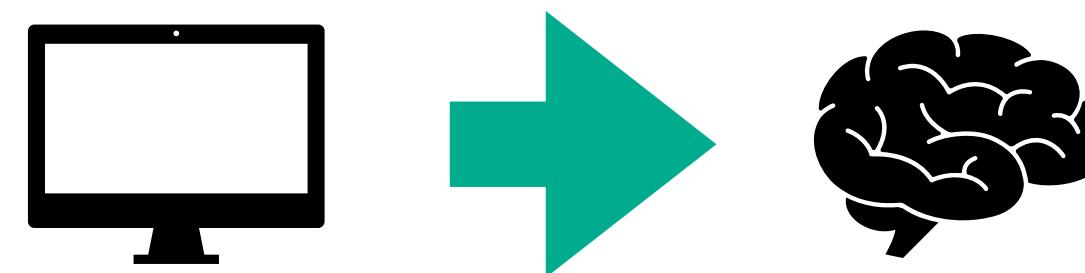
Natural Language Processing

Computer science + Linguistics + Machine Learning

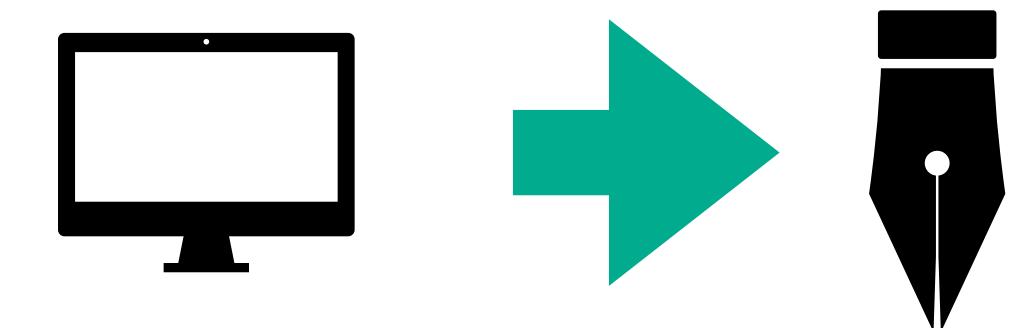
Speech Recognition



Natural Language
Understanding



Natural Language
Generation



AI Tools using NLP

Easy (mostly solved)	Intermediate (good progress)	Hard (still hard)
Spell and Grammar checking	Information retrieval	Question answering
Text categorization tasks	Sentiment analysis	Summarization
Named-entity recognition tasks	Machine translation	Dialogue system
	Information extraction	

** Many of these tools are not available for Tamil (some are not yet matured even for English)

NLP: Techniques

Syntactic analysis (Syntax): Parsing the language with rules of formal grammar

Parse

```
(ROOT
  (S
    (NP
      (NP (NNP Surgical) (NN resection) (NNS specimens))
      (PP (IN of)
        (NP
          (NP (CD 85) (JJ invasive) (JJ ductal) (NNS carcinomas))
          (PP (IN of)
            (NP
              (NP (CD 85) (NNS women))
              (SBAR
                (WHNP (WP who))
                (S
                  (VP (VBD had)
                    (VP (VBN undergone)
                      (NP (CD 3D) (NN ultrasound)))))))))))
      (VP (VBD were)
        (VP (VBN included)))
    (. .)))
```

Semantic analysis (Semantics): Process of understanding the meaning and interpretation of words, signs and sentence structure

- Synonymy: fall & autumn
- Hypernymy & hyponymy (is a): animal & dog
- Meronymy (part of): finger & hand
- Homonymy: fall (verb & season)
- Antonymy: big & small

Sentences that are syntactically correct need not be semantically correct

NLP: Pipeline

Preprocess texts to a common format using different techniques

Standardisation

Sentence: **The Sun@ Rises iN tHE EaST1!**

i. Case normalisation:
lowercase

the sun@ rises in the east1!

ii. Punctuation removal

the sun@ rises in the east1

iii. Remove unwanted symbols

the sun rises in the east

iv. Stop word removal

sun rises in east

Process of splitting the text into smaller units

Tokenization

Engrams - unigrams, bigrams, etc.

Sentence: **The lion is the king of the jungle.**

Unigram:

The, lion, is, the, king, of, the, jungle.

Bigram:

The lion, lion is, is the, the king, king of, of the, the jungle.

Process of converting token into its base form (morpheme)

Normalization

Token can have the structure,
<prefix> <morpheme> <suffix>

Sentence: **Antisocialist**

Anti social ist

Rule-based process that removes inflectional forms from a token (stem)

Stemming

Stem need not be a meaningful word.

Sentence: **"His teams are not winning"**

Stem:

"hi", "team", "are", "not", "winn"

Step-by-step process of removing inflectional forms from a token (lemma)

Lemmatization

Using vocabulary, word structure, part of speech tags, and grammar relations (lemma). Lemmas are root words.

Example:

Running, Run, Ran >> Run

NLP: Pipeline

Vectorization

Maps words or phrases from vocabulary to a corresponding vector of real numbers (semantics)

Methods:

- Bag of words (BoW)
- Tf-idf (Term Frequency – Inverse Document Frequency)
- Word embeddings (Word2Vec)

Raw Text

A dog in heat needs more than shade	Bag of words vector
	Dog 0
	need 2
	Cat 1
	than 0
	it 1
	heat 2
	needs 0

$$w_{x,y} = tf_{x,y} \times \log \left(\frac{N}{df_x} \right)$$

TF-IDF

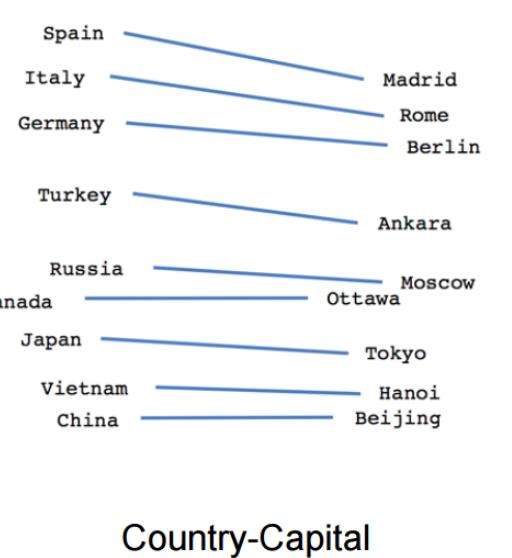
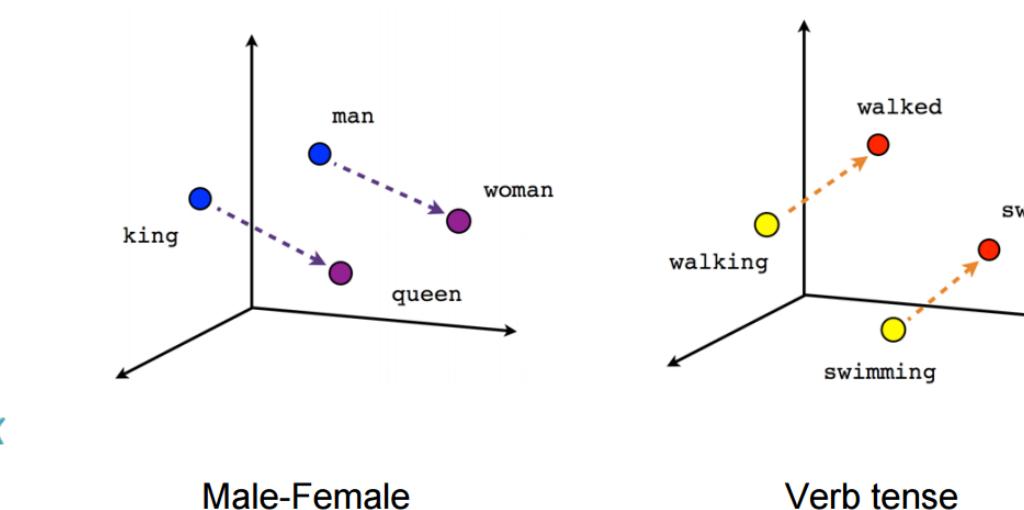
Term x within document y

$tf_{x,y}$ = frequency of x in y
 df_x = number of documents containing x
 N = total number of documents

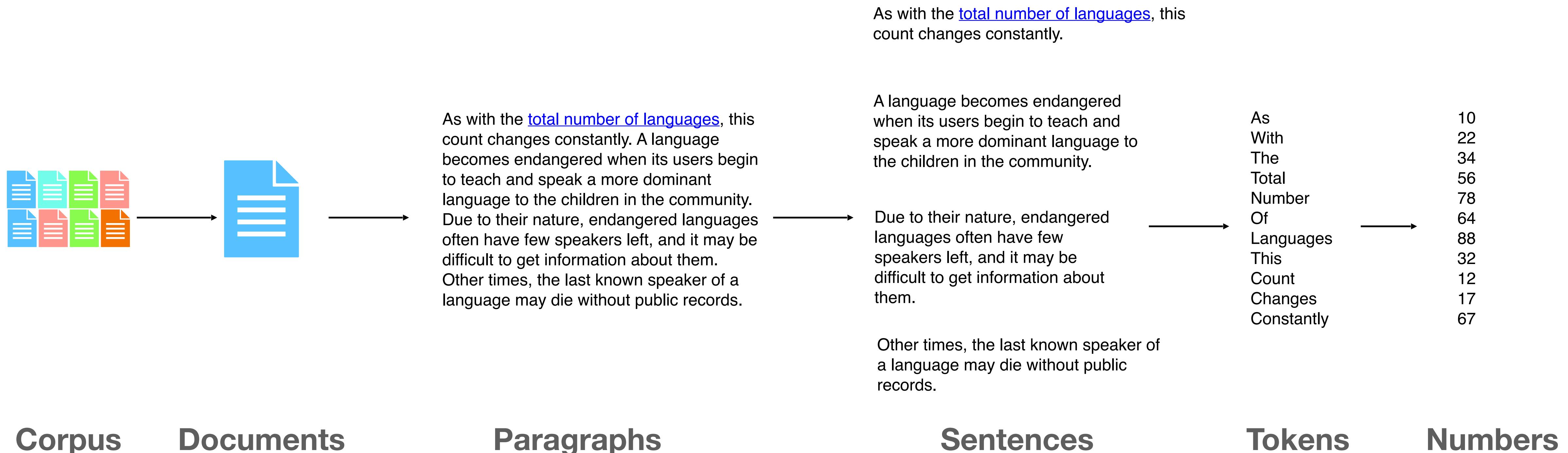
Bag of words (BoW)

TF-IDF

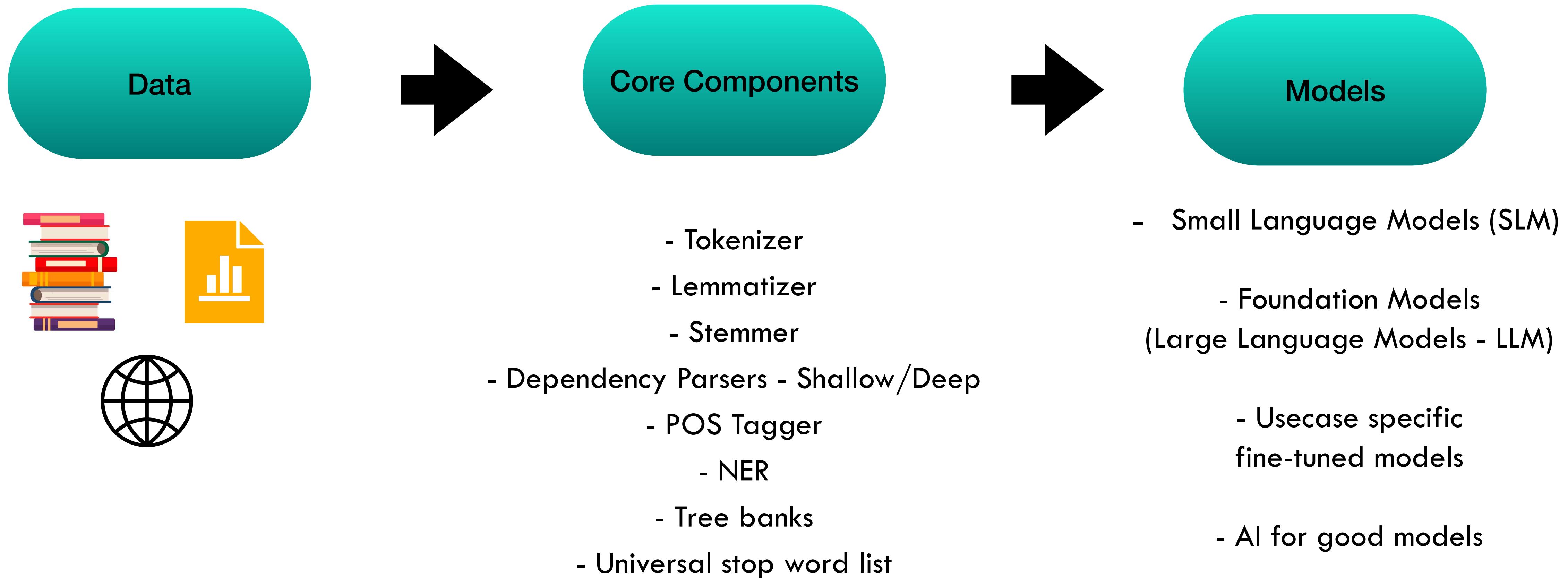
Word embeddings



Data: NLP Pipeline



Building Tamil AI - Necessities



Data: Curation Challenges

Stop looking down on the data curation process!



Time consuming

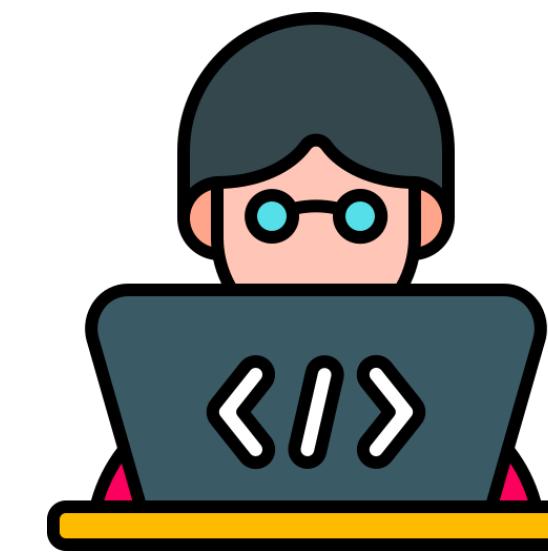


Labour intensive



Identify incentives

- Monetary reward
- Co-authorship
- Swags



Builder
mindset

vs



Consumer
mindset

Government can support open source initiatives!

Data: Curation Framework

Vidhai: How to contribute?

Creation



VIDHAI

An AITamilNadu Initiative

- Manually curate high quality datasets depending on the NLP task.
- Scrape content from books, websites, online forums, [Project Madurai](#) etc

Sign up here: <https://aitamilnadu.org/>

Validation

- Scrapped or machine-generated data by linguists, and NLP experts.
- Existing multi-lingual datasets that contains Tamil and perform translation and quality check ([AI4Bharat](#), [Bhashini](#), [Aya by Cohere for AI](#))

Data: Curation Framework

Vidhai: Who should get involved?

Community

- Linguists
- NLP experts
- ML Researchers
- Research Engineers
- Students
- Public



An AITamilNadu Initiative

Sign up here: <https://aitamilnadu.org/>

Government /
Institutions /
Organizations

- Monetary support for contributors
- Infrastructure grants
- Provide access to data resources (textbooks)
- Regulations on usage
- Fostering communities
- Promote open source initiatives
- Taking technology to the masses

Core Components

Either improve existing libraries or build from scratch!

iNLTK

- Tokenization
- Word embeddings
- Sentence similarity
- Text completion

Indic NLP Library

- Normalization
- Transliteration
- Phonetic analysis
- Syllabification

StanfordNLP/Stanza

- Lemmatization
- Parts-of-Speech (POS)
- Named Entity Recognition (NER)
- Dependency parsing

** All these libraries support many Indian languages including Tamil but the quality of the output for Tamil still needs to be improved.

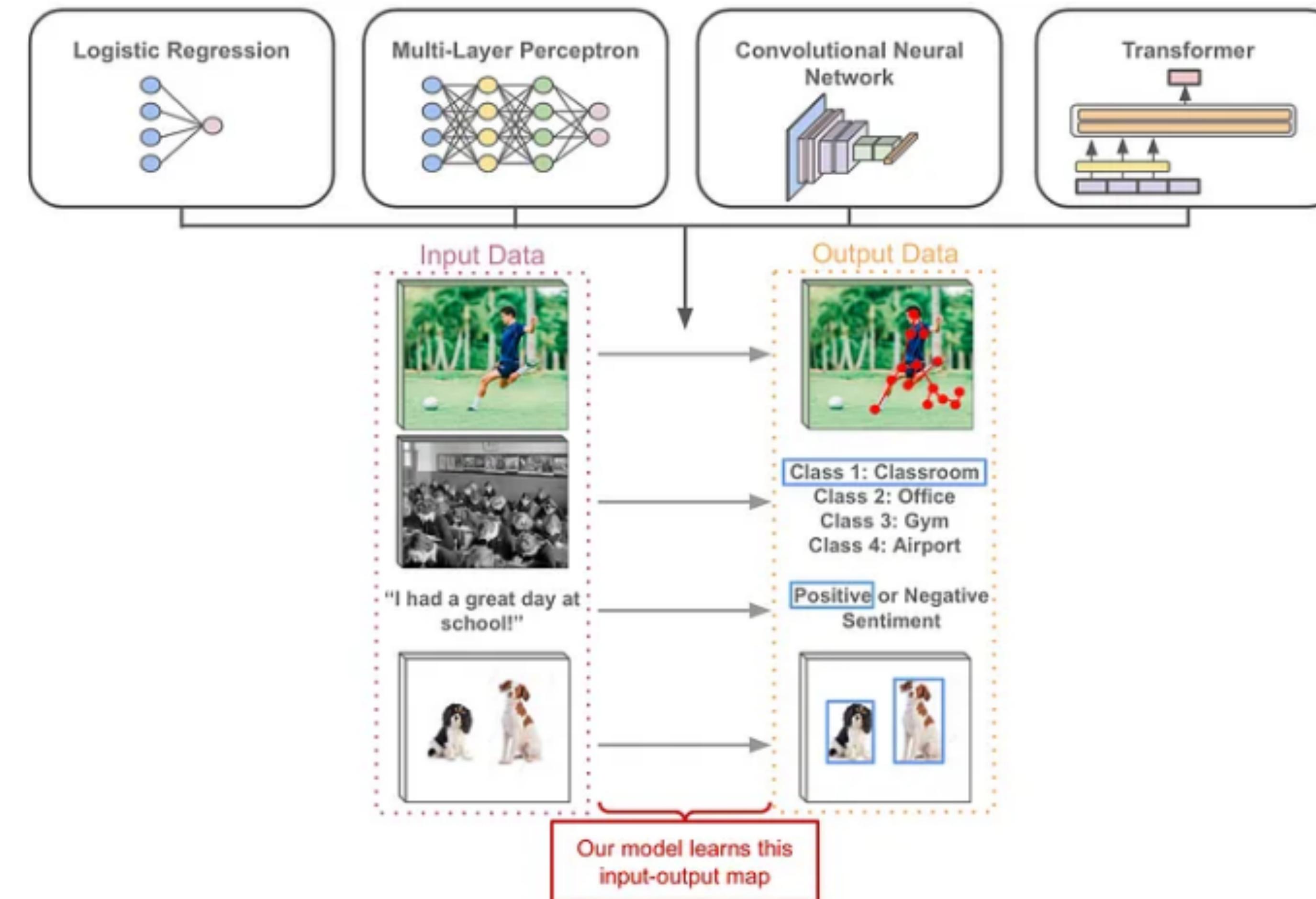
ThamizhiMorph - Morphological Parser

ThamizhiUDP - A Tamil Universal Dependency Parser

ThamizhiPOSt - A POS tagger for Tamil

Exclusive Tamil components

Models: Traditional Approach



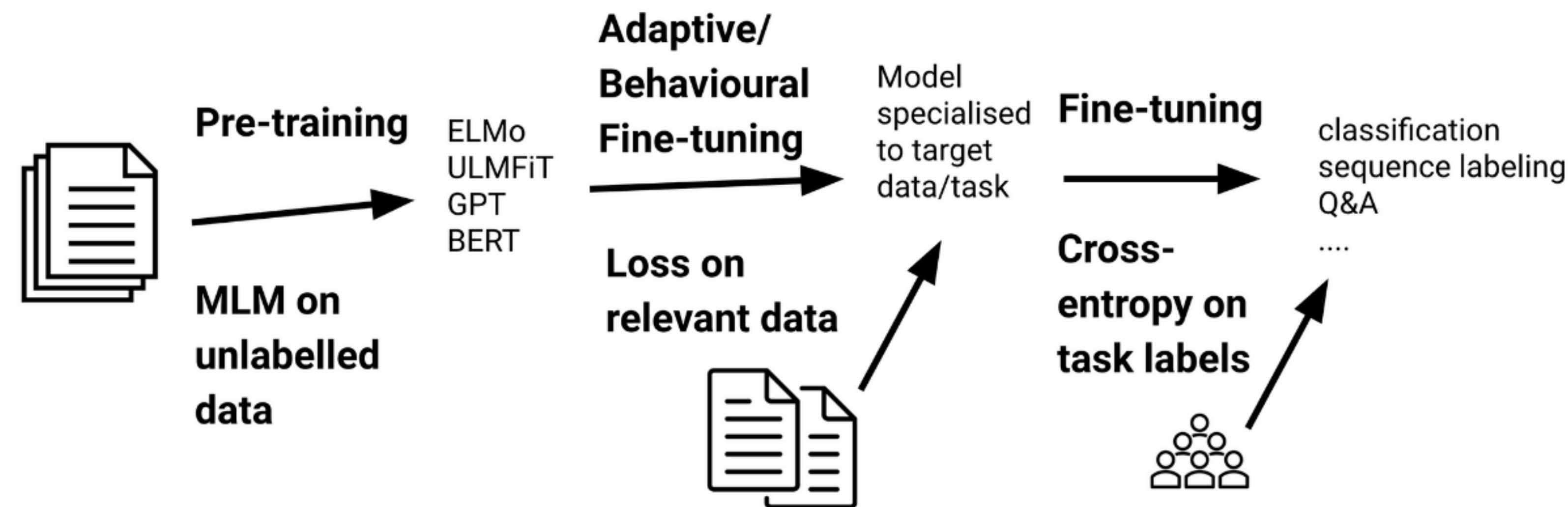
Models: Transfer Learning

Step 1: Pre-training

- Use large amounts of generic data and train on a specific objective function.
- Unlabelled data is used to train on the language modelling objective like MLM.

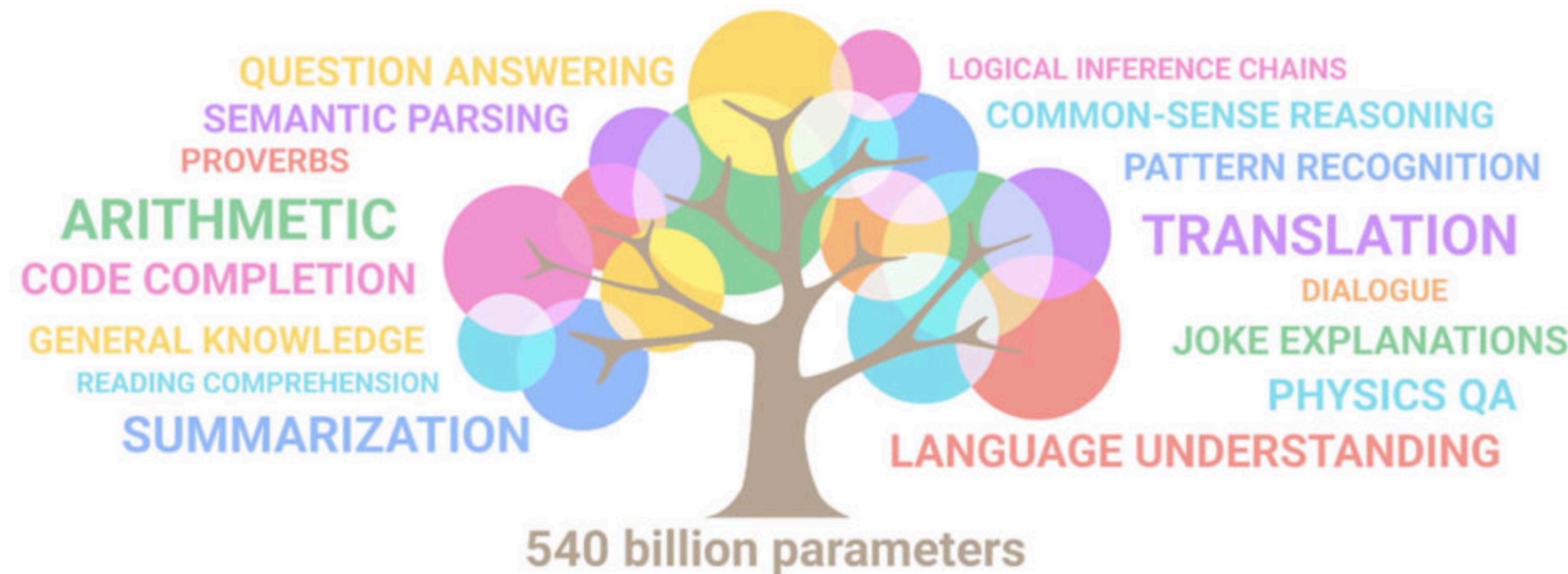
Step 2: Fine-tuning

- Fine-tuning is done using task-specific objective function.
- Labelled data is used to fine-tune model on the downstream tasks.

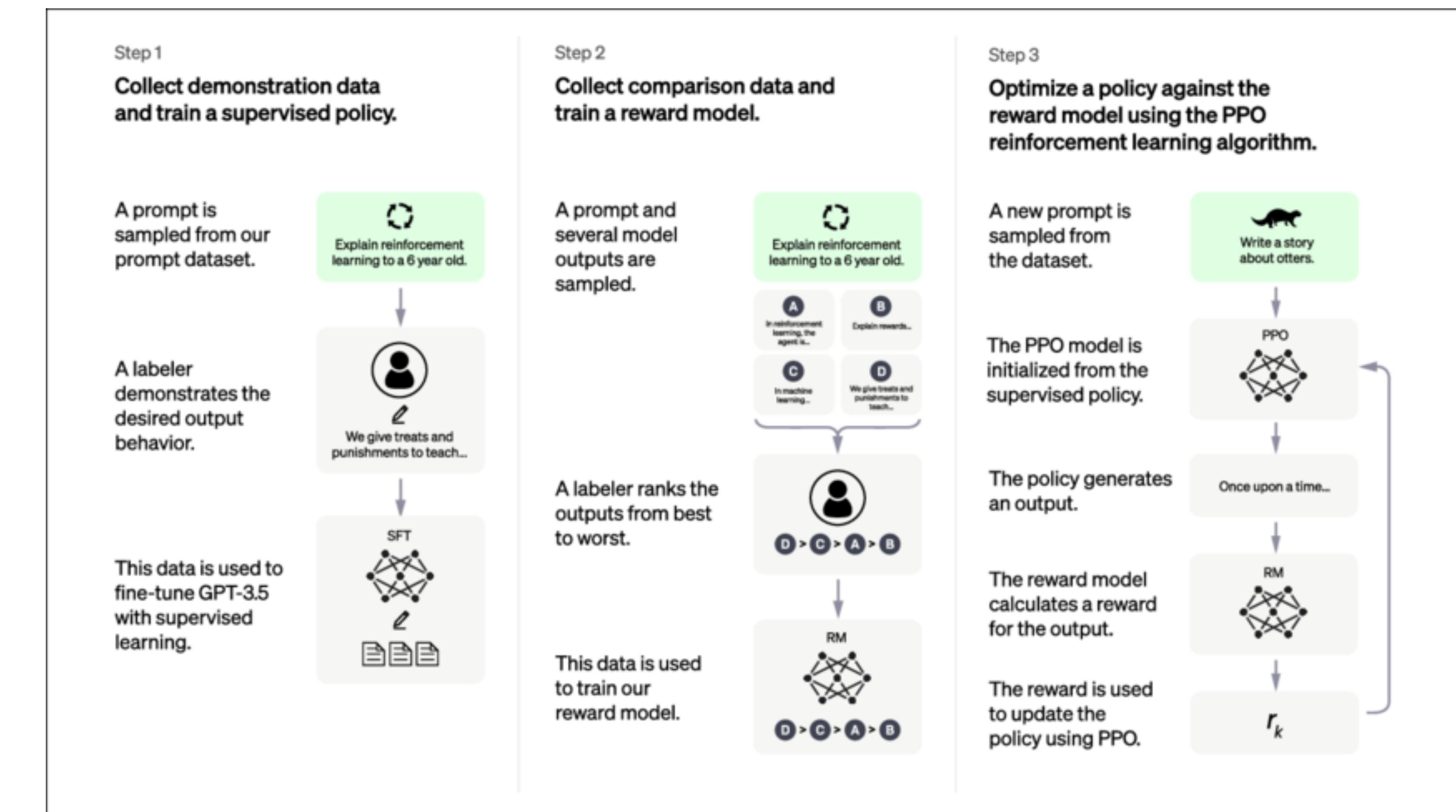
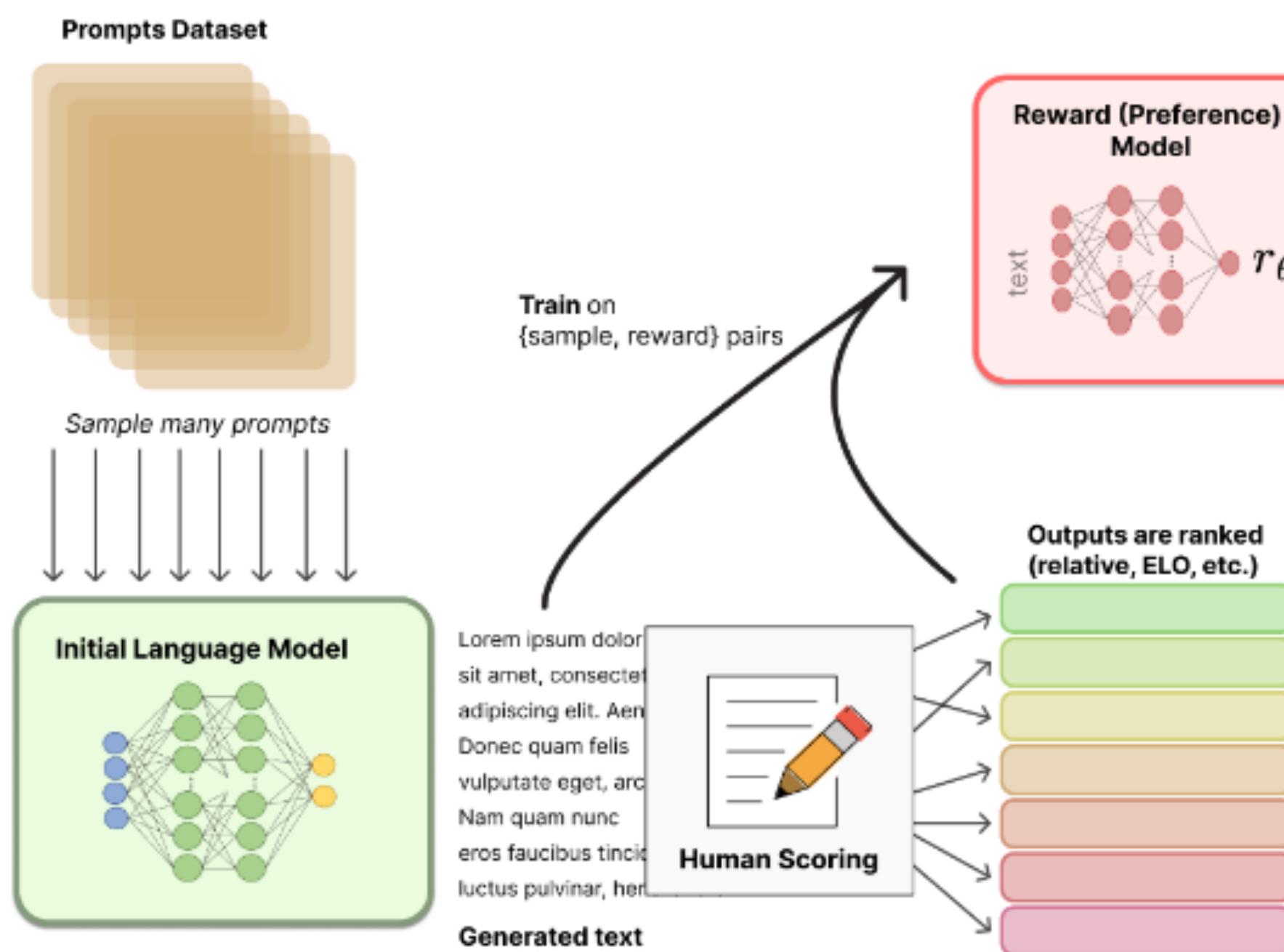


Foundation Models / Large Language Models

A foundation model is a large AI model pre-trained on a vast quantity of unlabelled data that was "designed to be adapted" (or fine-tuned) to a wide range of downstream tasks, such as sentiment analysis, image captioning, and object recognition. Prompt Engineering is used to interact with the model.



LLM Training: RLHF



<https://huggingface.co/blog/rlhf>

The ChatGPT training process. The figure is from OpenAI (2022a).

Prompt Engineering

- A prompt is a short piece of text that is given to the large language model as input, and it can be used to control the output of the model in many ways.
- Designing this prompt efficiently is called prompt engineering.

A prompt contains any of the following elements:

Instruction - a specific task or instruction you want the model to perform

Context - external information or additional context that can steer the model to better responses

Input Data - the input or question that we are interested to find a response for

Output Indicator - the type or format of the output.

Models for Tamil

GPT-2 Tamil (Abinaya Mahendiran)

This screenshot shows the Hugging Face model card for 'abinayam/gpt-2-tamil'. The card includes sections for Model card, Downloads last month (1,630), Safetensors (Model size: 137M params, Tensor type: F32 · U8), Hosted inference API (Text Generation, Examples), and a detailed description of the model's purpose and setup. It also features a command-line interface for running 'pip install -r requirements.txt'.

GPT2-Tamil

This repository is created as part of the Flax/Jax community week by Huggingface. The aim of this project is to pretrain a language model using GPT-2 specifically for Tamil language.

Setup:

To setup the project, run the following command,

```
pip install -r requirements.txt
```

Model:

Pretrained model on Tamil language using a causal language modeling (CLM) objective.

Tamillama (Raju Kandasamy)

This screenshot shows the Hugging Face model card for 'RajuKandasamy/tamillama_tiny_30m'. The card includes sections for Model card, Downloads last month (460), Hosted inference API (Text Generation, Examples), and a detailed description of the model's purpose and usage. It also features a large text area with Tamil text and a summary in English.

Tamillama_Tiny: A 30M tiny LLaMA model trained to tell stories in Tamil

TL;DR:

This is an experimental model inspired by the paper <https://arxiv.org/abs/2305.07759> - How Small Can Language Models Be and Still Speak Coherent English?

Extended the same concept for Tamil. A 30M parameter LLaMA architecture model that outputs coherent Tamil is presented here.

Additional experimentation which is included in the model:

1. This is a multilanguage model as it can output both English and Tamil stories.
2. The model also does translation of stories from English to tamil and vice versa. To see the translation feature, set the max_new_tokens > 512.
3. Translation of original stories from the tinystories dataset was done using IndicTrans

For now, this is a toy model for researchers, students and LLM enthusiasts to play with the linguistic capability of the model.

Weights Release, License and Usage

We release the weights in two formats: Hugging Face transformers format and GGML format to use with CTransformers or LLaMA.cpp.

Models for Tamil

GPT-2 Tamil (Alagu Prakalya)

Lagstill/GPT-2-Tamil

Model description

GPT2-Tamil is a GPT-2 transformer model fine Tuned on a large corpus of Tamil data in a self-supervised fashion. This means it was pretrained on the raw texts only, with no humans labelling them in any way with an automatic process to generate inputs and labels from those texts. More precisely, it was trained to guess the next word in sentences.

More precisely, inputs are sequences of continuous text of a certain length and the targets are the same sequence, shifted one token (word or piece of word) to the right. The model uses internally a mask-mechanism to make sure the predictions for the token i only uses the inputs from 1 to i but not the future tokens.

This way, the model learns an inner representation of the Tamil language that can then be used to extract features useful for downstream tasks.

Tamil-LLaMa (Abhinand Balachandran)

Tamil-LLaMA-7B-Instruct-GGUF (CPU Demo)

Tamil LLaMA 7B Instruct v0.1 GGUF format model. Running the Q5_KM Quantized version.

Running on free CPU hardware. Suggest duplicating this space to run without a queue.

Note: The inference is quite slow as it is running on CPU.

Chatbot

வணக்கம், நீங்கள் யார்?

AI உதவியாளராக, எனக்கு தனிப்பட்ட முறையில் பெயர் இல்லை. இருப்பினும், நான் உங்களுக்கு உதவும் பதிலளிக்கவும் வடிவமைக்கப்பட்டுள்ளேன். உங்கள் கேள்விகளுக்கு துல்லியமான மற்றும் பயனுள்ள பதில்களை வழங்குவதில் நான் எப்போதும் மகிழ்ச்சியடைகிறேன்.

Type a message...

Submit

Retry Undo Clear

Examples

வணக்கம், நீங்கள் யார்? நான் பெரிய பணக்காரன் இல்லை, லேட்டஸ்ட் iPhone-இல் நிறைய பணம் செலவழிக்க வேண்டுமா?

பொடியலை வரிசைப்படுத்த பைதான் செயல்பாட்டை ஏழுதவும். சீவப்பும் மஞ்சளங்கும் கலந்தால் என்ன நிர்மாக இருக்கும்? விரைவாக தூங்குவது எப்படி?

Additional Inputs

Limitations of Generative AI



Hallucinations

- Not enough data
- Data is noisy
- Lack of fact checking
- No constraints



Bias

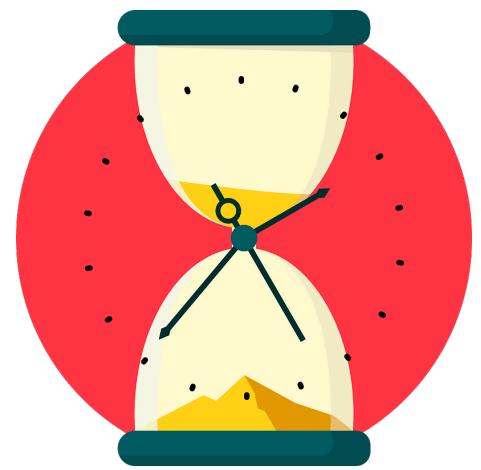
- Sampling bias
- Prejudice bias
- Confirmation bias
- Group attribution bias



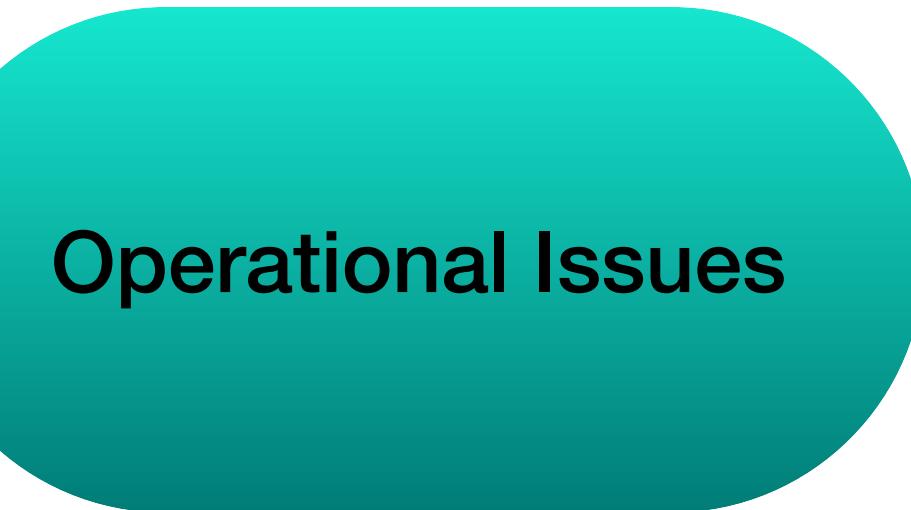
Ethical Regulation

- Guardrails
- Monitoring usage

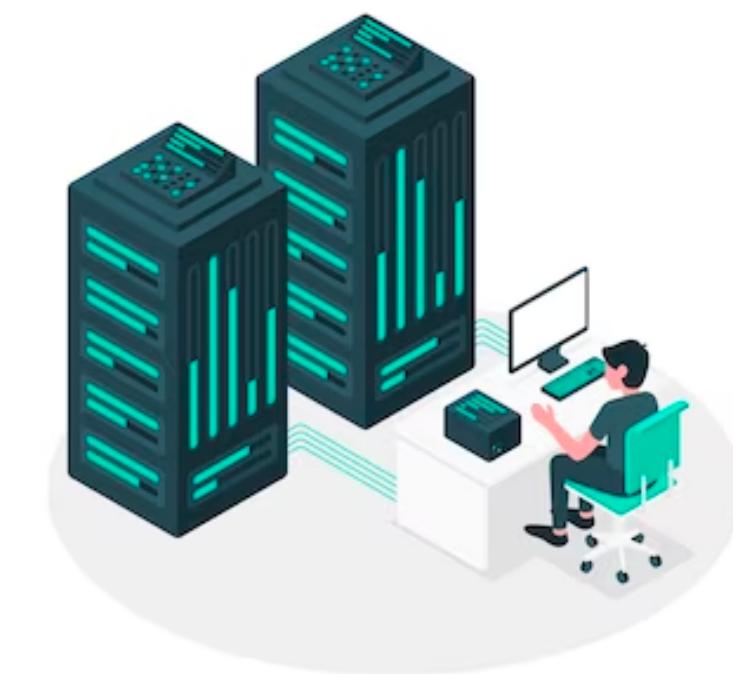
Limitations of Generative AI



Time



Cost



Infrastructure

Better Data >> More Data >> Clever Algorithms

Fine-tuned Models >> SLM >> LLM

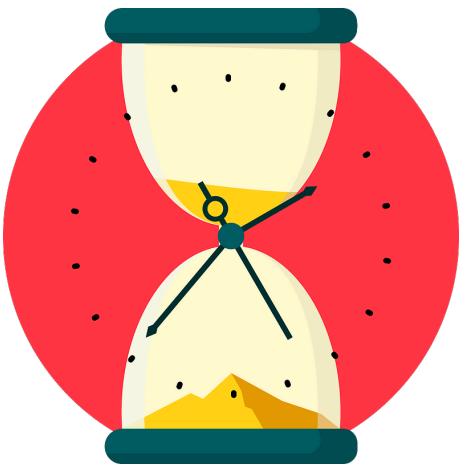
Gen AI: Research Directions

Open challenges in LLM research

1. Reduce and measure hallucinations
2. Optimize context length and context construction
3. Incorporate other data modalities
4. Make LLMs faster and cheaper
5. Design a new model architecture
6. Develop GPU alternatives
7. Make agents usable
8. Improve learning from human preference
9. Improve the efficiency of the chat interface
10. Build LLMs for non-English languages

Hardest is to build LLMs for non-English languages!

Role of the Community



Contribute your time



Impart technical/linguistic knowledge



Open Source



Provide monetary support
(CSR)



Support your local community
initiatives

Thank You :)

Questions?