# 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

@freakynut

https://medium.com/@abinayamahendiran

https://www.linkedin.com/in/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

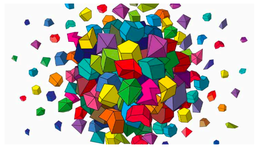
Limitations of Gen AI

Gen AI: Research Directions

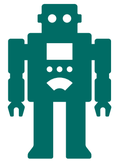
Role of the community

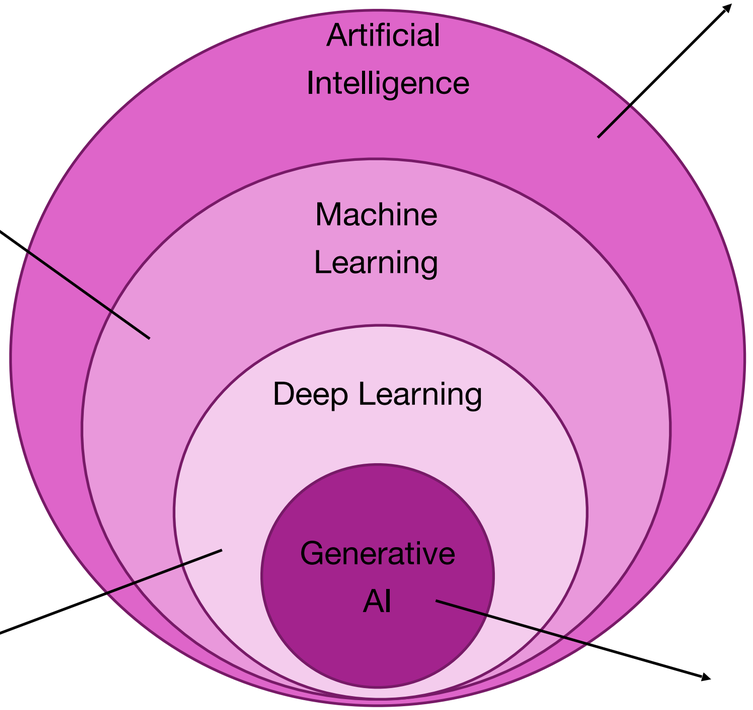
# Types of AI

Leverage historical data and make computers learn patterns without explicit programming



Make agents that can simulate human behaviour

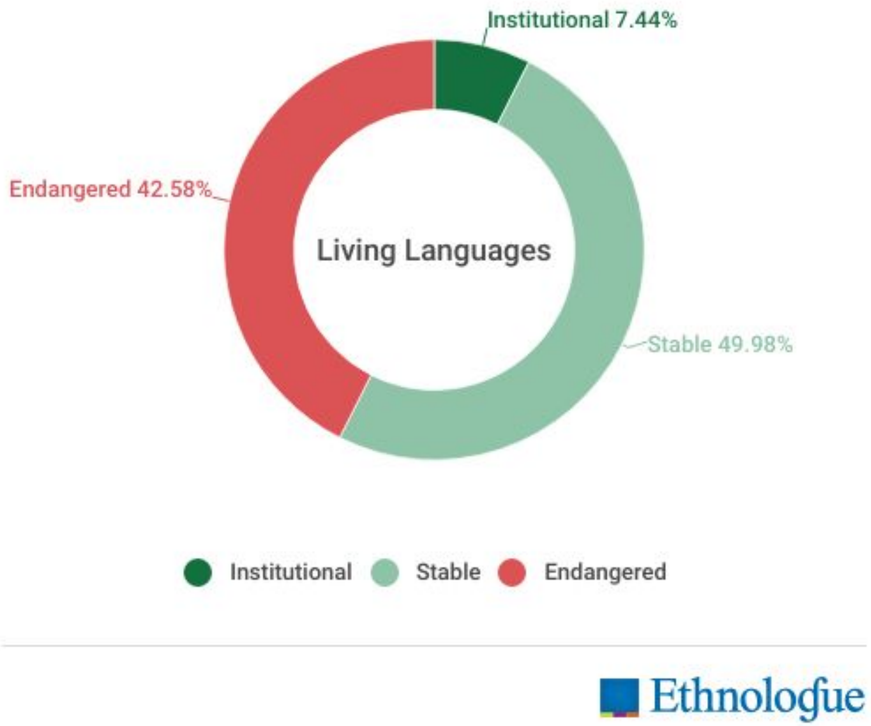





Mimic human brain through neural networks

Create new content based on the learned patterns from existing data

# Why Tamil AI?



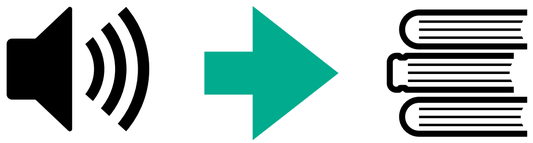

| சங்கம் | கால இடைவெளி | கவிஞர்களின் எண்ணிக்கை | இராச்சியம் [9] | புத்தகங்கள் [9] |
|---|---|---|---|---|
| முதலில் | 4440 ஆண்டுகள் [9] | 549 [9] | பாண்டியா | புத்தகங்கள் எதுவும் பிழைக்கவில்லை |
| இரண்டாவது | 3700 ஆண்டுகள் [9] | 1700 [9] | பாண்டியா | *தொல்காப்பியம்* (ஆசிரியர் – தொல்காப்பியர்) |
| மூன்றாவது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

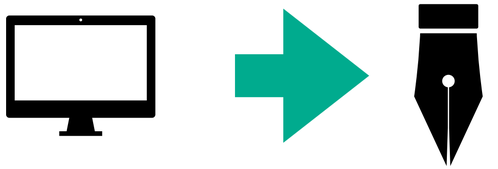
Computer science + Linguistics + Machine Learning

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**

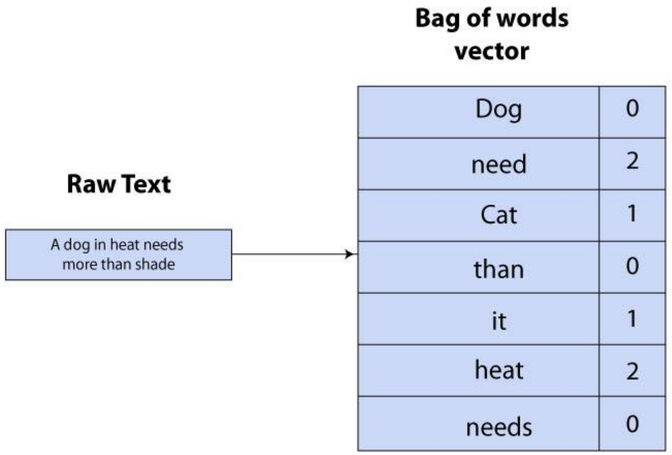
# NLP: Pipeline

**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)




Bag of words (BoW)

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

**TF-IDF**

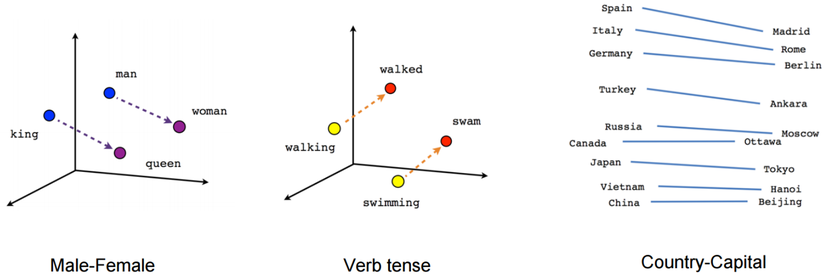
Term $x$ within document $y$

$tf_{x,y}$ = frequency of $x$ in $y$

$df_x$ = number of documents containing $x$

N = total number of documents

TF-IDF




Male-Female | Verb tense | Country-Capital

Word embeddings

# Data: NLP Pipeline

As with the total number of languages, this count changes constantly. A language becomes endangered when its users begin to teach and speak a more dominant language to the children in the community. Due to their nature, endangered languages often have few speakers left, and it may be difficult to get information about them. Other times, the last known speaker of a language may die without public records.

As with the total number of languages, this count changes constantly.

A language becomes endangered when its users begin to teach and speak a more dominant language to the children in the community.

Due to their nature, endangered languages often have few speakers left, and it may be difficult to get information about them.

Other times, the last known speaker of a language may die without public records.

| | |
|---|---|
| As | 10 |
| With | 22 |
| The | 34 |
| Total | 56 |
| Number | 78 |
| Of | 64 |
| Languages | 88 |
| This | 32 |
| Count | 12 |
| Changes | 17 |
| Constantly | 67 |

**Corpus** **Documents** **Paragraphs** **Sentences** **Tokens** **Numbers**

# Building Tamil AI - Necessities

Data

Core Components

- Tokenizer
- Lemmatizer
- Stemmer
- Dependency Parsers - Shallow/Deep
- POS Tagger
- NER
- Tree banks
- Universal stop word list

Models

- Small Language Models (SLM)

- Foundation Models
(Large Language Models - LLM)

- Usecase specific
fine-tuned models

- AI for good models

# 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**

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

VIDHAI

An AITamilNadu Initiative

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

**Validation**

- Scraped 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

VIDHAI

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.

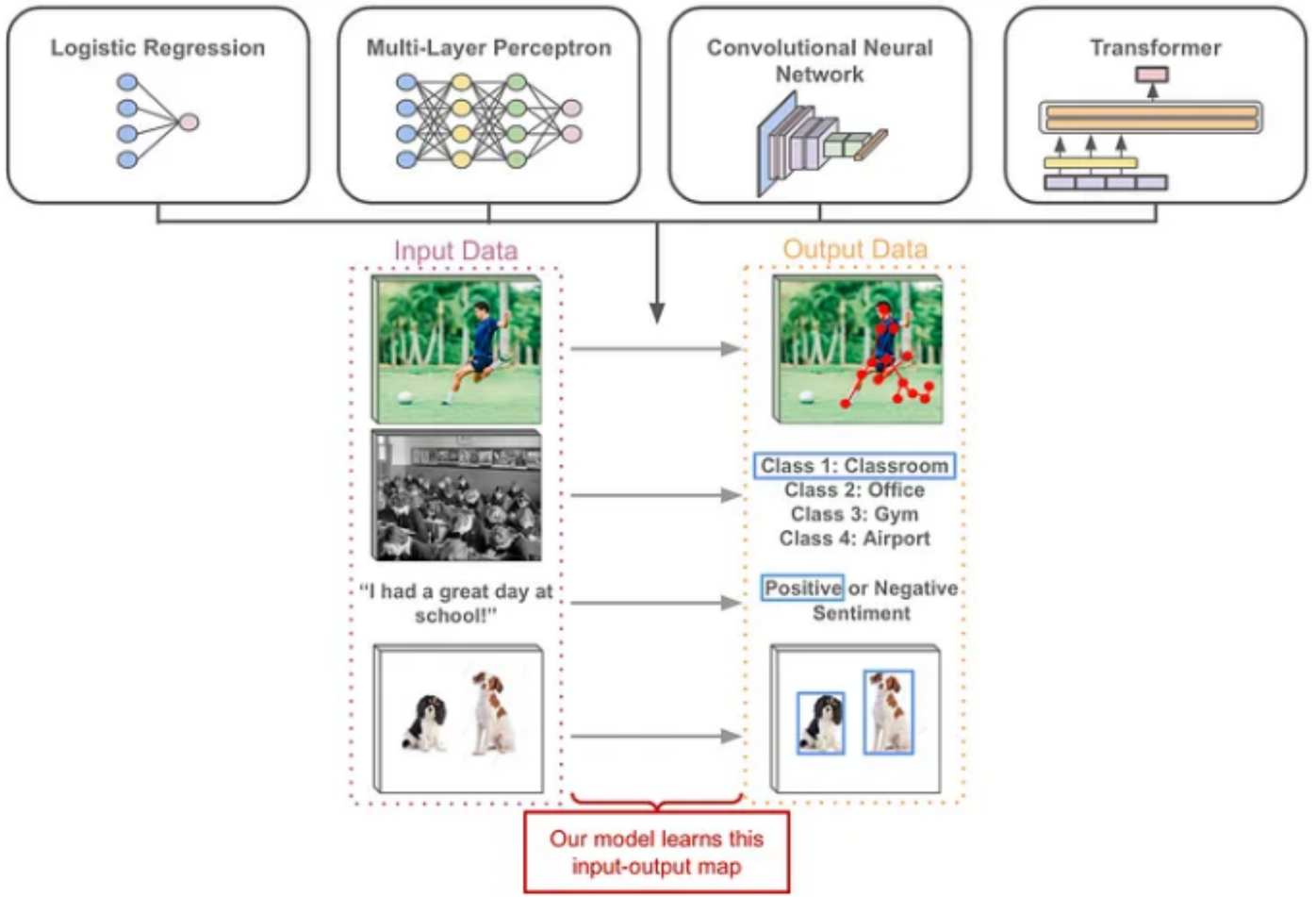
**ThamizhiMorph - Morphological Parser**

**ThamizhiUDP - A Tamil Universal Dependency Parser**

**ThamizhiPOSt - A POS tagger for Tamil**

Exclusive Tamil components

# Models: Traditional Approach




https://towardsdatascience.com/language-models-gpt-and-gpt-2-8bdb9867c50a

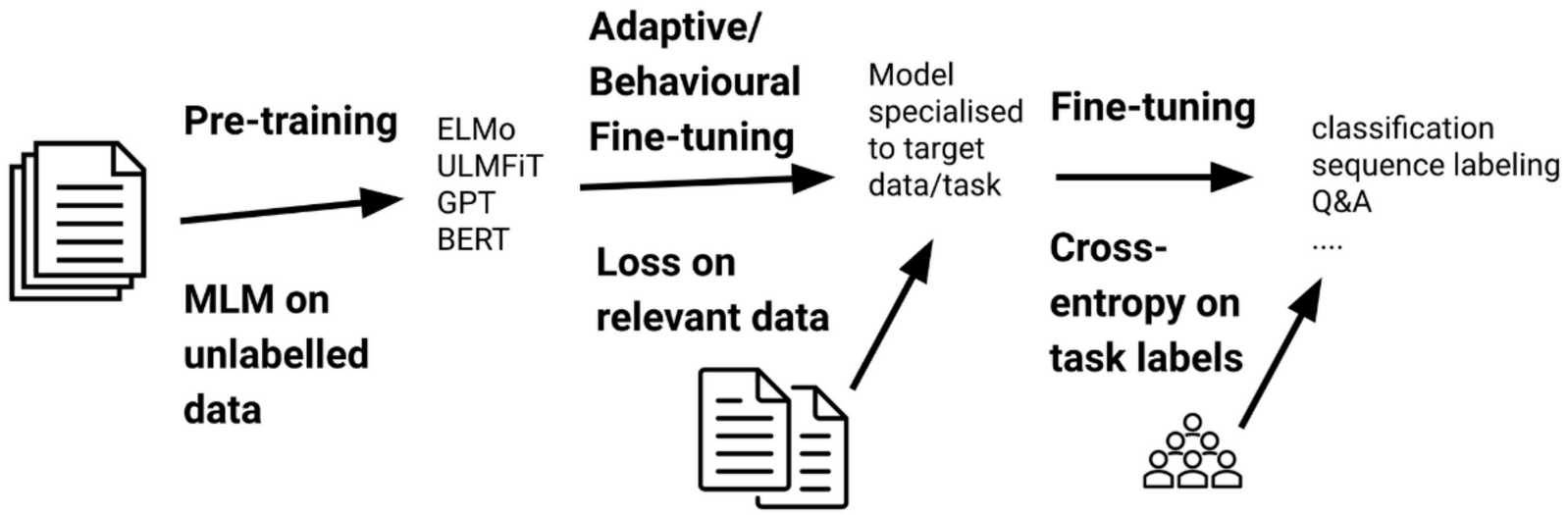
# Models: Transfer Learning

**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**

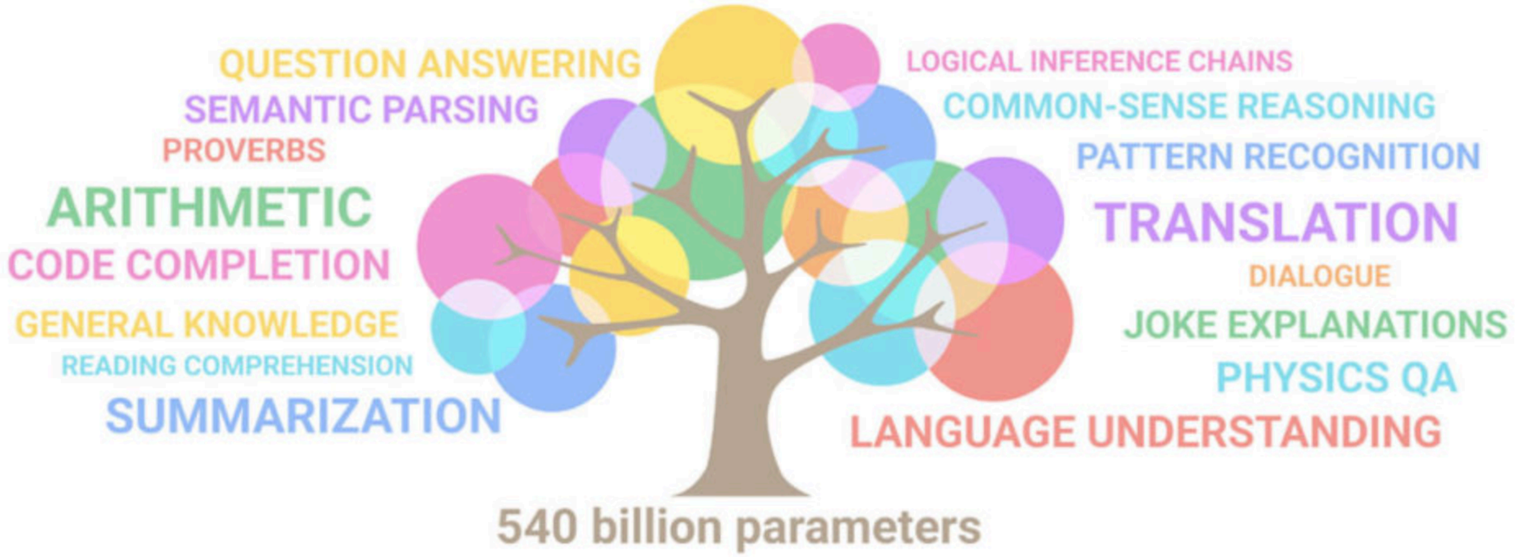
- Fine-tuning is done using task-specific objective function.
- Labelled data is used to fine-tune model on the downstream tasks.



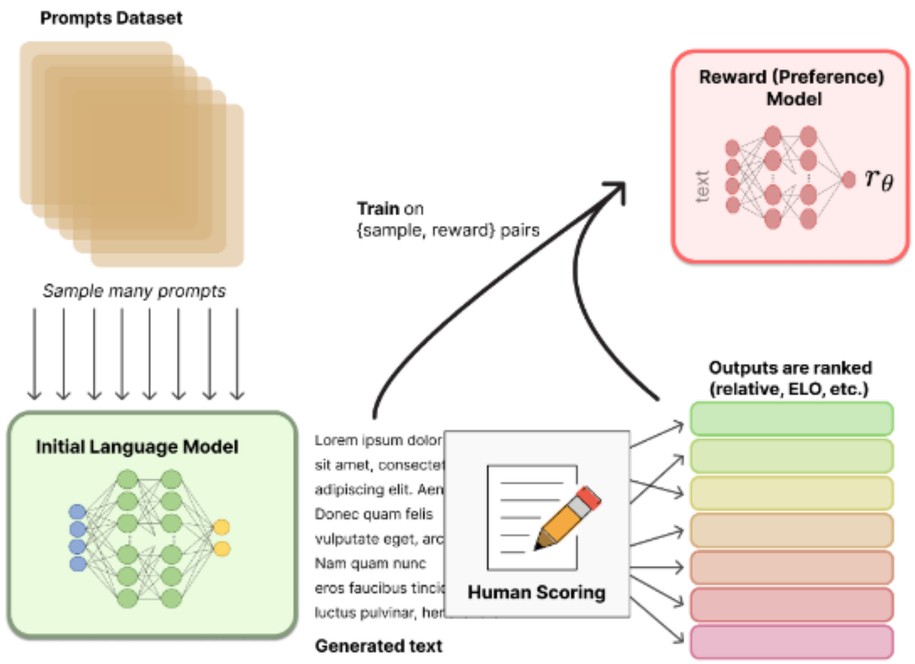

https://www.ruder.io/recent-advances-lm-fine-tuning/

# Foundation Models / Large Language Models

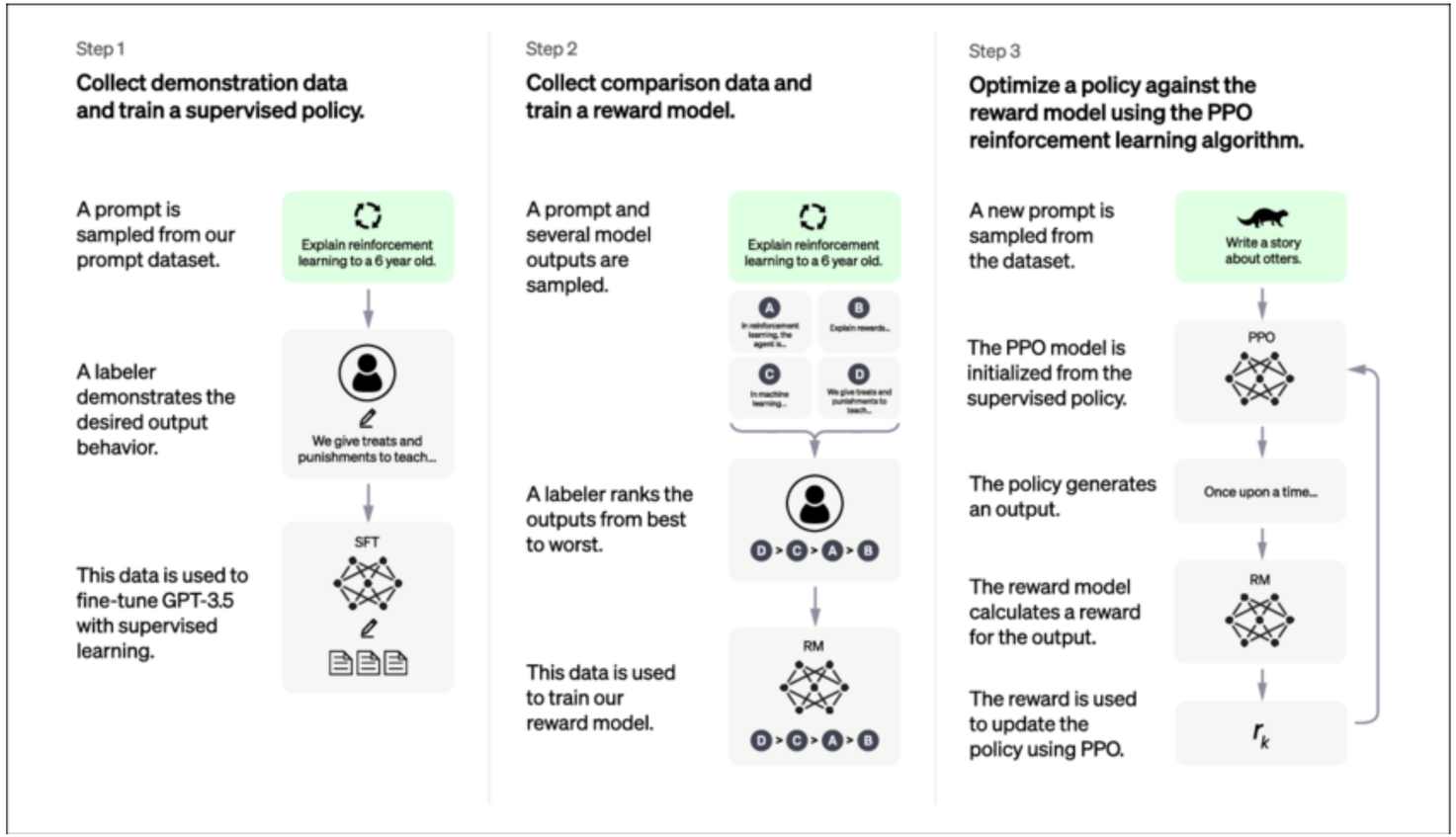
A foundation model is a large AI model pre-trained on a vast quantity of unlabelled data that was "designed to be adapted" (or fine-tuned) to a wide range of downstream tasks, such as sentiment analysis, image captioning, and object recognition. Prompt Engineering is used to interact with the model.



# 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

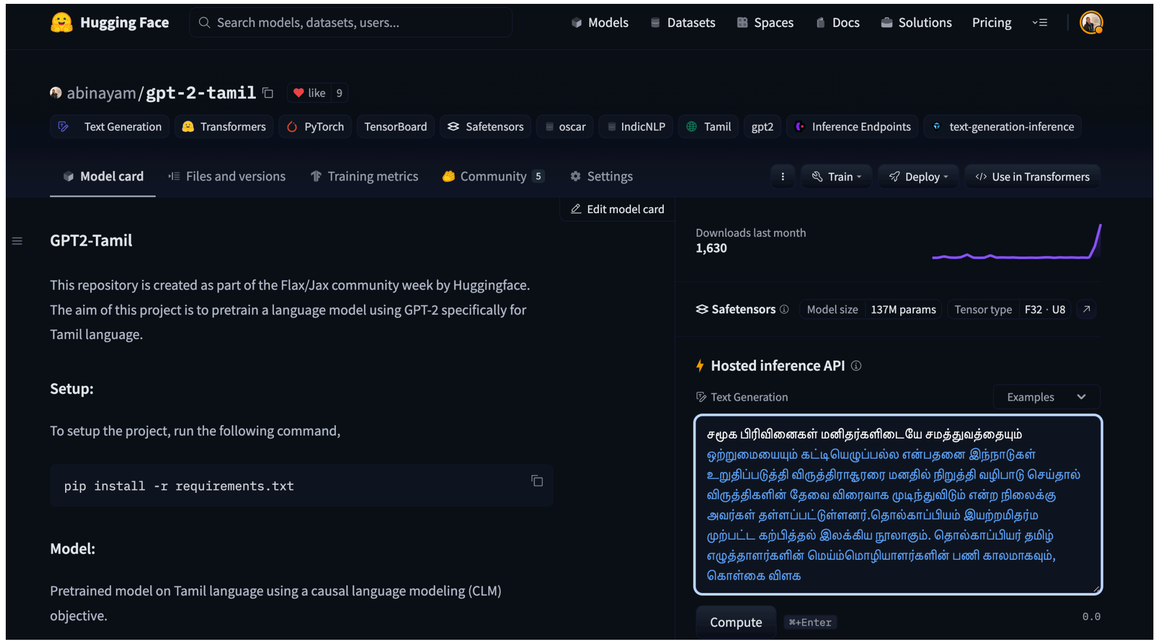
**Input Data** - the input or question that we are interested to find a response for

**Output Indicator** - the type or format of the output.

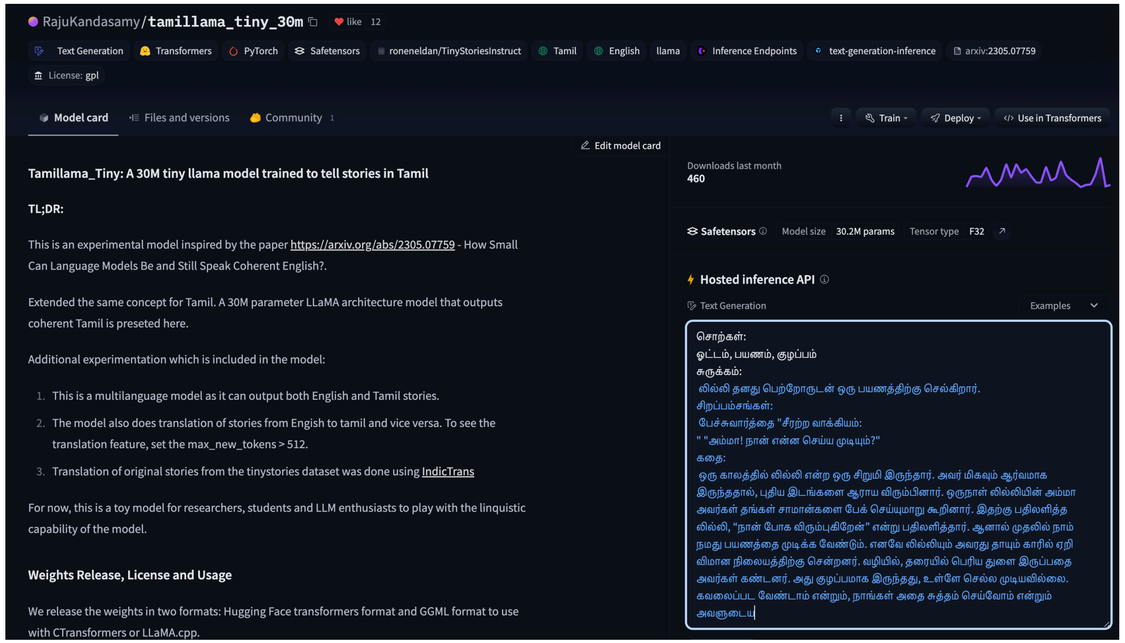
https://www.promptingguide.ai/introduction/elements

# Models for Tamil

## GPT-2 Tamil (Abinaya Mahendiran)

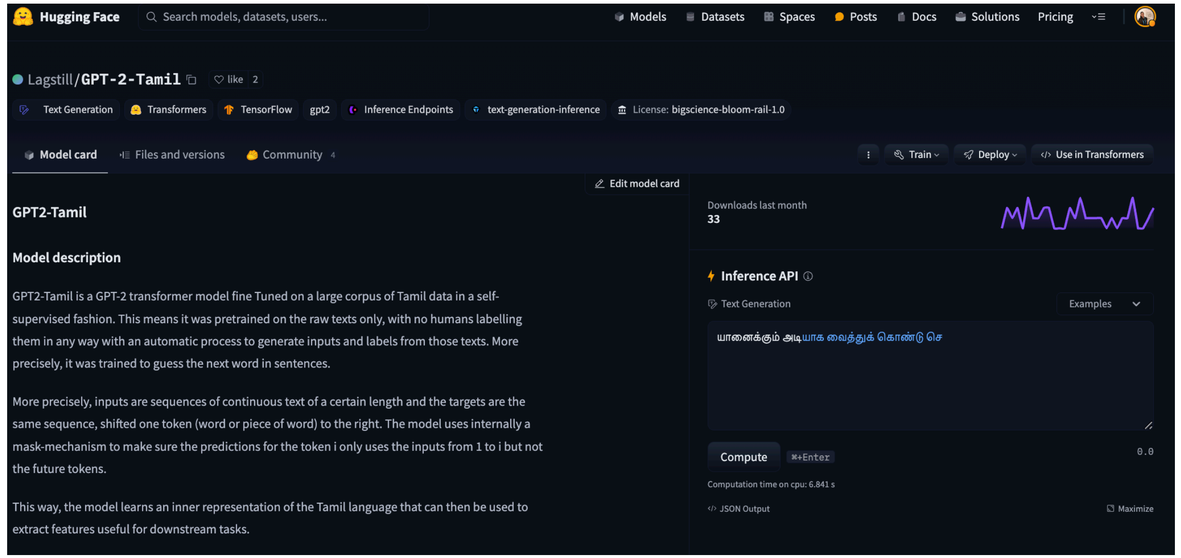



## Tamillama (Raju Kandasamy)

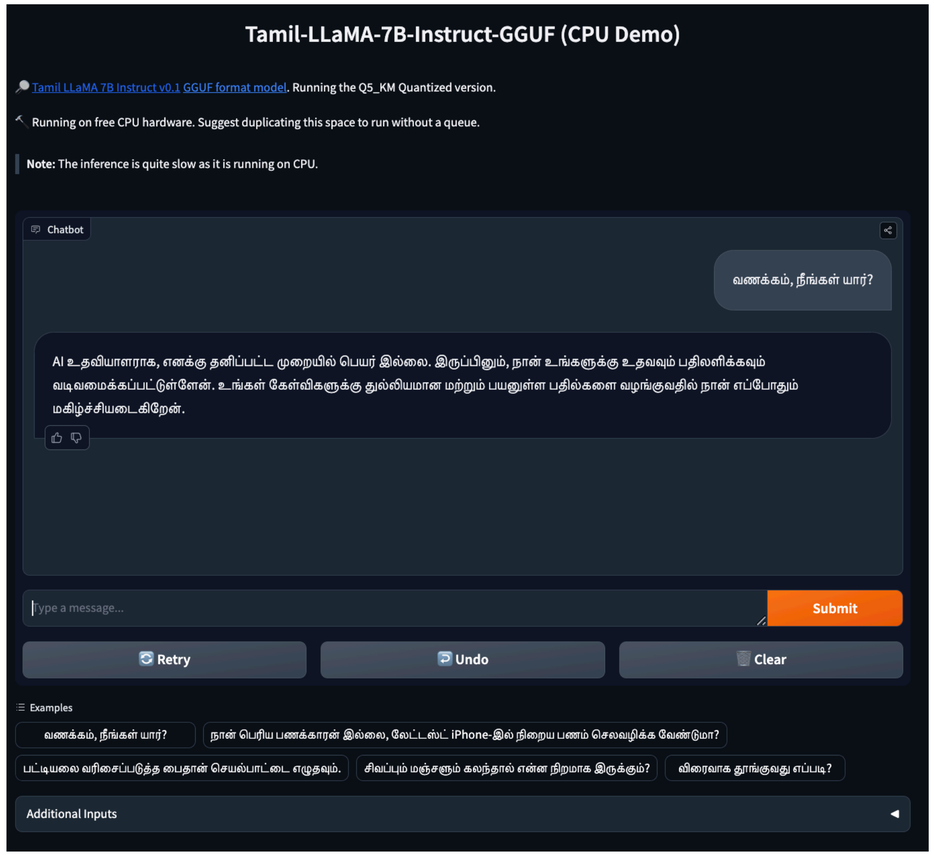


# Models for Tamil

## GPT-2 Tamil (Alagu Pragalya)




## Tamil-LLaMa (Abhinand Balachandran)



# Limitations of Generative AI

Data/Governanace Issues

## 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

Operational Issues

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!**

https://huyenchip.com/2023/08/16/llm-research-open-challenges.html

# Role of the Community

Contribute your time

Provide monetary support (CSR)

Impart technical/linguistic knowledge

Open Source

Support your local community initiatives

Thank You :)

Questions?