CS626: Speech, NLP and Web Introduction

Pushpak Bhattacharyya
Computer Science and Engineering Department
IIT Bombay

Week of 29th July, 2024

This and the follow-up course

- CS626 (Autumn)- *NLP, Speech, and Web*:
  - Concept Building, Task Understanding, Technique Building

- CS772 (Spring)- *Deep Learning for Natural Language Processing*
  - NLP concepts covered in CS626 will see their realization on Deep Neural Nets

Course Logistics

Lecture Venue

- LA001
- Ground floor of lecture hall complex
- Opp CSE building

Course website

Course Website:

https://www.cse.iitb.ac.in/~cs626/2024/

For information on NLP research, visit CFILT website.

https://www.cfilt.iitb.ac.in/

https://www.cse.iitb.ac.in/~pb/

Moodle

Login to Moodle with LDAP credentials.

- Select the course CS626

All course related notifications will be notified via Moodle also.

Evaluation Scheme (tentative)

- 50%: Reading, Thinking, Comprehending
  - Quizzes (10)
  - Midsem (15)
  - Endsem (25)
- 50%: Doing things, Hands on
  - Assignments (25%)
  - Reading ONE paper and doing a preliminary implementation of the same (25%)

– Quiz every last Thursday of the month

Journals and Conferences

- Journals: Computational Linguistics, Natural Language Engineering, Journal of Machine Learning Research (JMLR), Neural Computation, IEEE Transactions on Neural Networks, TACL

- Conferences: ACL, EMNLP, NAACL, EACL, AACL, NeuriPS, ICML, UAI, AIStat

Useful NLP, ML, DL libraries

- NLTK
- Scikit-Learn
- Pytorch
- Tensorflow (Keras)
- Huggingface
- Spacy
- Stanford Core NLP
- CFILT tools and resources

NLP=Language+Computation

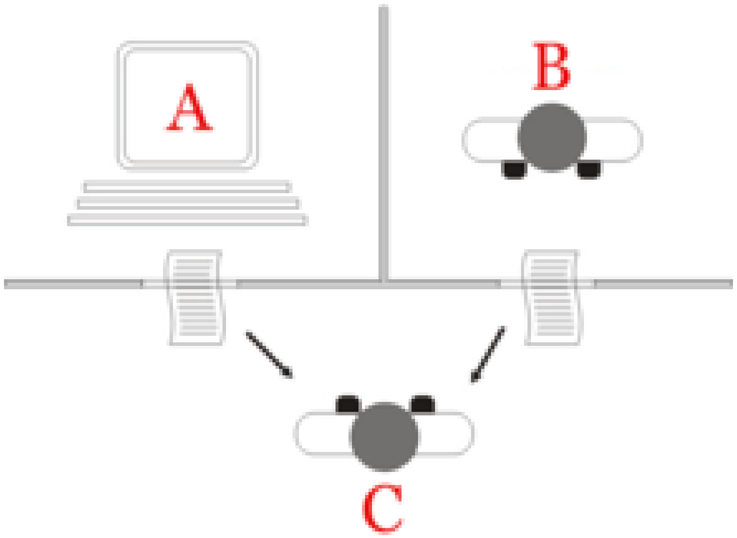
A useful video

https://www.ted.com/talks/patricia_kuhl_the_linguistic_genius_of_babies?geo=hi&subtitle=en

Two Foundations

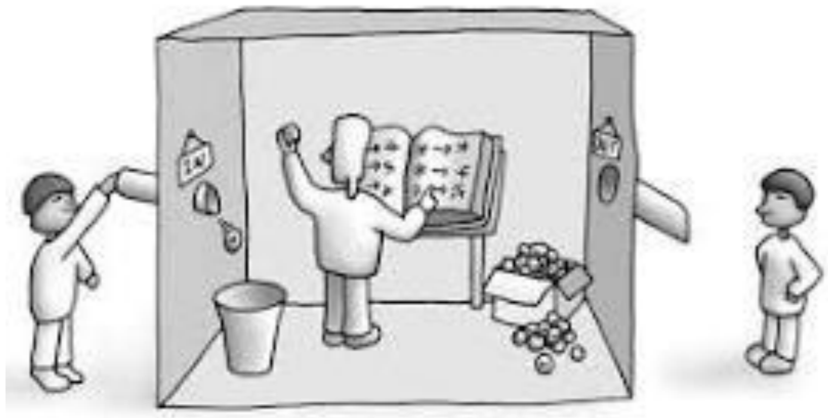
Turing Test *(wikipedia)*

- The **Turing test**, originally called the **imitation game** by Alan Turing in 1950
- Test of a machine's ability to exhibit intelligent behavior
- Equivalent to, or **indistinguishable** from, that of a human



The "standard interpretation" of the Turing test, in which player C, the interrogator, is given the task of trying to determine which player – A or B – is a computer and which is a human. The interrogator is limited to using the responses to written questions to make the determination

Searl’s Chinese Room Experiment



- A computer program cannot have a "mind", "understanding", or "consciousness", regardless of how intelligently or human-like the program may make the computer behave.

- Philosopher John Searle presented the argument in his paper "Minds, Brains, and Programs", published in *Behavioral and Brain Sciences* in 1980.

- **A human being sits in the room and does exactly**

I gave this test (1/2)

Read this story and examine everything is right about this story: "once a ship was going from England to France. After a few days of journey, it was caught in a violent storm. The captain tried a lot to steer the ship clear from the eye of the storm, but the ship sped to a cliff and dashed against it. All those on board died and only the captain survived. With a binocular, he could spot a spot of jungle very far. With great difficulty, he took his articles near the jungle and after going there fell unconscious. He regained consciousness after 3 days. After becoming conscious, the captain felt very hungry. He fortunately found some potatoes amongst his articles. (cntd.)

I gave this test (2/2)

He managed to get a fire ignited by striking two pieces of stone. There were some dry pieces of wood which he lighted and boiled the potatoes on the fire and appeased his hunger. When he had slept for some time, he woke up and found some shining objects on the sand where the fire was burnt. He kept these objects with him. After several days, he spied a ship far away with the binocular. The ship rescued him from the island. The captain took those shining objects with him to England. Later they found interesting properties of these shining objects. These objects were transparent, but could be used to make containers. The material was given the name "glass". This is how glass was invented."

ChatGPT response (1/2)

There are several issues with this story, indicating that it is not entirely accurate.

Firstly, glass was not invented by a shipwrecked captain as described in the story. The history of glassmaking goes back thousands of years and involves the contributions of many different cultures.

Secondly, the story implies that glass is made from potatoes, which is completely false. Glass is made from a mixture of silica, soda ash, and limestone, among other materials.

ChatGPT response (2/2)

Thirdly, the story suggests that the captain discovered glass by chance, simply by finding some shiny objects on the sand. However, the process of making glass is complex and requires specialized knowledge and equipment.

Finally, the story is written in a somewhat simplistic and unconvincing manner, with several inconsistencies and unrealistic details, such as the captain managing to survive on a deserted island with only a few potatoes and no other resources for several days.

In summary, while the story may be entertaining, it is not a reliable source of information about the invention of glass.

Regenerate response (1/2)

There are a few issues with the story:

- The story describes the invention of glass, but glass was actually first produced by the ancient Mesopotamians around 3500 BCE, long before the time period in which the story is set.
- The story suggests that the captain single-handedly invented glass by discovering it on a deserted island. However, the invention of glass was a complex and gradual process that involved experimentation and refinement by many people over many centuries. (Cntd.)

Regenerate response (2/2)

Cntd.

- The story does not provide any specific details about the captain's journey or the location of the island, which makes it difficult to verify the accuracy of the events described.
- In conclusion, while the story may be an interesting work of fiction, it is not an accurate portrayal of the invention of glass.

Introduction and Perspective

What is Language *(George Yule, “Study of Language”, 1998)*

- **Displacement** (Indicators that change with time and place: I saw him yesterday at the market; I will see him tomorrow in the *school*)
- **Arbitrariness** (name → Meaning; *water, chair*)
- **Productivity/creativity** (potentially infinite no. of sentences)
- **Cultural Transmission** (child acquires parent’s language)
- **Discreteness** (sound and meaning units separated)
- **Duality** (Surface structure, deep structure)

What is “language phenomenon”

- Sandhi (phonetic phonological transformation at morpheme boundaries)
  - *गुरु + आदेश = गुर्वादेश: guru + aadesh = gurvaadesh (The order from the teacher)*
- Fronting of verbs in question formation in German
  - *Trinkst du Tee? (Do you drink tea)*
- Semantic/Pragmatic Incongruity leading to Irony, Sarcasm
  - An irate unattended invitee leaving the party and being asked *“how did you enjoy the party?”* and replying *“O, I love being ignored”*

There are Rare language phenomena

- E.g., Sarcasm
  - 11% of Tweets are Sarcastic
  - *Out of that 17% is numerically sarcastic*

Modelling of language phenomena: softmax

$$\sigma(\bar{Z})_i = \frac{e^{Z_i}}{\sum_{j=1}^{K} e^{Z_j}}$$

- σ is the **softmax** function
- Z is the input vector of size K
- The RHS gives the ith component of the output vector
- Input to softmax and output of softmax are of the same dimension

Example

$$\bar{Z} = <1,\ 2,\ 3>$$

$$Z_1 = 1,\ Z_2 = 2,\ Z_3 = 3$$

$$e^1 = 2.72,\ e^2 = 7.39,\ e^3 = 20.09$$

$$\sigma(\bar{Z}) = < \frac{2.72}{2.72+7.39+20.09}, \frac{7.39}{2.72+7.39+20.09}, \frac{20.09}{2.72+7.39+20.09} >$$

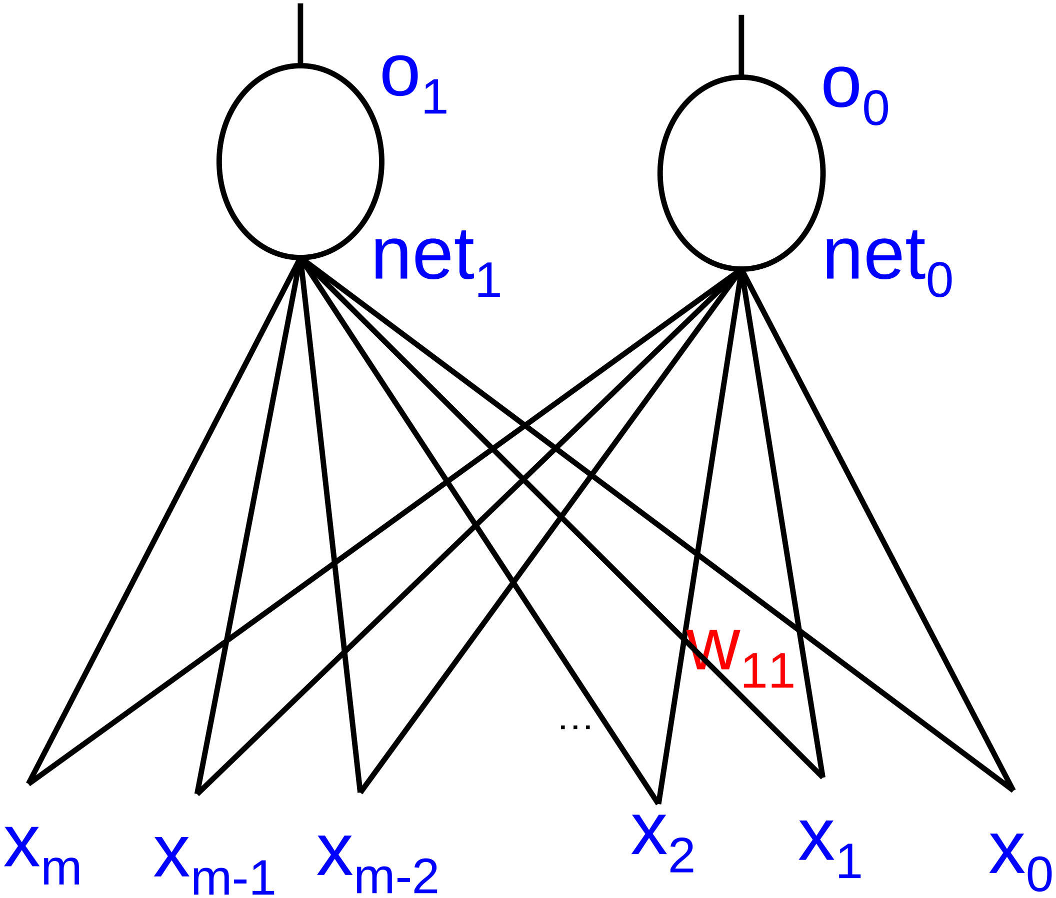
$$= < .09,\ 0.24,\ 0.67 >$$

Modelling of language phenomena: Cross Entropy

$$H(P,Q) = -\sum_{x=1,N} \sum_{k=1,C} P(x,k) \log_2 Q(x,k)$$

x varies over *N* data instances, *c* varies over *C* classes
P is target distribution; *Q* is observed distribution

Multiple neurons in the output layer: softmax+*cross entropy* loss (1/2): illustrated with 2 neurons and single training data point



$$O = < o_1, o_0 >$$

$$NET = < net_1, net_0 >$$

$$o_1 = \frac{e^{net_1}}{e^{net_1} + e^{net_0}},\ o_0 = \frac{e^{net_0}}{e^{net_1} + e^{net_0}}$$

$$\frac{\partial O}{\partial NET} = \begin{bmatrix} \frac{\partial o_0}{\partial net_0} & \frac{\partial o_1}{\partial net_0} \\ \frac{\partial o_0}{\partial net_1} & \frac{\partial o_1}{\partial net_1} \end{bmatrix}$$

$$= \begin{bmatrix} o_0(1-o_0) & -o_0 o_1 \\ -o_1 o_0 & o_1(1-o_1) \end{bmatrix}$$

Softmax and Cross Entropy (2/2)

$$L = -t_1 \log o_1 - t_0 \log o_0$$

$$o_1 = \frac{e^{net_1}}{e^{net_1} + e^{net_0}}, \; o_0 = \frac{e^{net_0}}{e^{net_1} + e^{net_0}}$$

$$\frac{\partial L}{\partial w_{11}} = -\frac{t_1}{o_1}\frac{\partial o_1}{\partial w_{11}} - -\frac{t_0}{o_0}\frac{\partial o_0}{\partial w_{11}}$$

$$\frac{\partial o_1}{\partial w_{11}} = \frac{\partial o_1}{\partial net_1} . \frac{\partial net_1}{\partial w_{11}} + \frac{\partial o_1}{\partial net_0} . \frac{\partial net_0}{\partial w_{11}} = o_1(1-o_1)x_1 + 0$$

$$\frac{\partial o_0}{\partial w_{11}} = \frac{\partial o_0}{\partial net_1} . \frac{\partial net_1}{\partial w_{11}} + \frac{\partial o_0}{\partial net_0} . \frac{\partial net_0}{\partial w_{11}} = -o_1 o_0 x_1 + 0$$

$$\Rightarrow \frac{\partial L}{\partial w_{11}} = -t_1(1-o_1)x_1 + t_0 o_1 x_1 = -t_1(1-o_1)x_1 + (1-t_1)o_1 x_1$$

$$= [-t_1 + t_1 o_1 + o_1 - t_1 o_1]x_1 = -(t_1 - o_1)x_1$$

$$\Delta w_{11} = -\eta \frac{\partial E}{\partial w_{11}} = \eta(t_1 - o_1)x_1$$

Nature of NLP

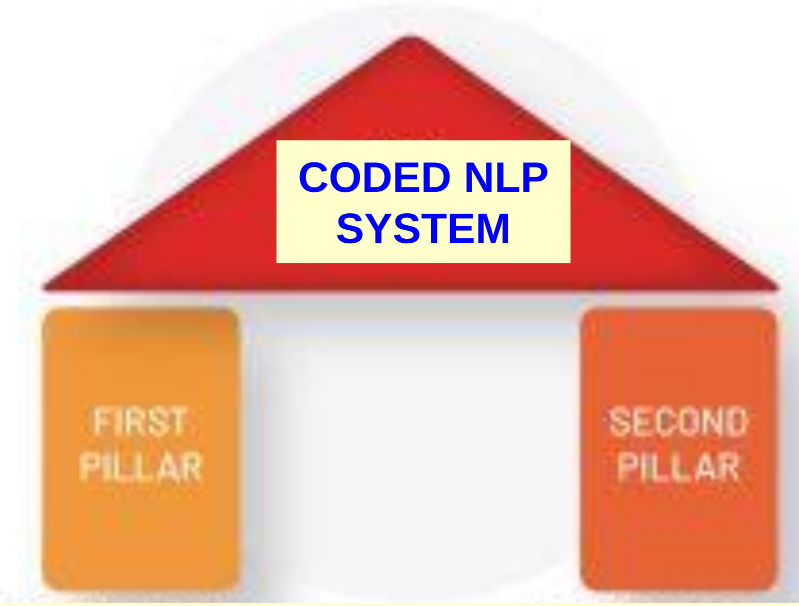
Natural Language Processing

Art, science and technique of making computers understand and generate language

3 Generations of NLP

- Gen1- Rule based NLP is also called Model Driven NLP
- Gen2- Statistical ML based NLP (*Hidden Markov Model, Support Vector Machine*)
- Gen3- Neural (Deep Learning) based NLP
- Gen3.5- LLM and GenAI

Two Pillars of NLP



LINGUISTICS + PROBABILITY
= NLP

Linguistics → Language Insight: 1st pillar of NLP

- **BEFORE** the model operates:
  - *Preparing the training data; Deciding on the annotation scheme (e.g., POS tag labels)*
  - *Guidelines for annotation (e.g., in ‘golf club’, ‘golf’ is an adjectival, i.e., noun functioning as adjective)*
- **AFTER** the model has operated
  - *Interpretation of model output (e.g., “I love being ignored” labeled as ‘sarcastic’, incongruity of ‘love’ and ‘ignore’)*
  - *Error Analysis (MT of Dravidian languages has low BLEU score due to morph complexity)*

Probability → Data Insight: 2nd pillar of NLP

- **LAW OF LARGE NUMBERS** operates
  - *NN follows JJ 80% of the time: P(NN|JJ)=0.8*
  - *Credibility increases as data increases*
- **CENTRAL LIMIT THEOREM** operates
  - *Beyond a threshold (~30) on the amount of data, we can use normal distribution approximation, whatever the underlying distribution to estimate the parameters*
  - *Can use hypothesis testing to ascertain the robustness of generalization and also measure statistically the PROGRESS (aka, improvement over baseline)*

Main challenge is Ambiguity!: an extreme example

“Buffalo buffaloes Buffalo buffaloes buffalo buffalo Buffalo buffaloes”

Prompt to chatGPT: what do you understand by the above sentence

chatGPT response

- Buffalo (noun): Refers to the city of Buffalo, which is in the state of New York in the United States.
- buffaloes (verb): Means to bully, confuse, or baffle.
- Buffalo (noun): Refers to the animal, specifically the American bison.

describes a scenario where bison from Buffalo are being intimidated or confused by other bison from the same place, who, in turn, are intimidating or confusing other bison. It's a playful sentence that plays on the multiple meanings of the word "Buffalo" as a place

CORRECT!!

Another ambiguity (ToI, 29jul24)

Language Modelling

What does a Model do?

- **The only thing Models** do is **PREDICT!**
- E.g.,

$$s = ut + \frac{1}{2}at^2$$

- *S:distance; u: initial velocity; a:acceleration; t: time*
- Now, given <*u, a, t*> we can compute (predict!) *s*

What does a Language Model do? (1/2)

- **A. Predicts** “next” language objects
  - E.g.-1, Answer to Question: *what is the capital of India* → *Delhi*
  - E.g.-2, Summary of input text: *Delhi is a large city. The city has large number of people residing and passing through. Large businesses are transacted here. The cultural activities are numerous.* → *Delhi is a large, populous and busy city*

What does a Language Model do? (2/2)

- **B. (application of A) Predicts** properties of language objects, called classification of regression
  - E.g., sentiment analysis: *The movie was wonderful → +ve sentiment*

The foundation

- All predictions on language objects are FOUNDED on ONE prediction task
- **Predict** the next word given a sequence of words

$$P(W_N \mid W_{N-1} W_{N-2} \ldots W_1 W_0)$$

$$= \frac{\#(W_0 W_1 \ldots W_{N-1} W_N)}{\#(W_0 W_1 \ldots W_{N-1})}$$

What is Curse of Dimensionality

- As the number of dimensions (features) increases, the amount of data required to adequately sample the space grows exponentially
- E.g., No. of Boolean Functions: 2^{2^N}
- No. of Threshold functions: 2^{N^2}

Curse of Dimensionality and LM

- N-gram LM
- $P(W_N|W_0W_1W_2W_3\ldots W_{N-1})$
- What happens as *N* increases?

$$P(W_N \mid W_{N-1}W_{N-2}\ldots W_1W_0) = \frac{\#(W_0W_1\ldots W_{N-1}W_N)}{\#(W_0W_1\ldots W_{N-1})}$$

Curse of Dimensionality and LM, cntd.

- N-gram LM, equivalent to computing $P(W_0W_1W_2W_3\ldots W_{N-1}W_N)$
- How many such parameters need to be computed?
- Let $|V|$ be the vocab size
- Each position of the n-word string can be filled in $|V|$ ways
- Hence the number of parameters is $|V|^N$
- *Curse of dimensionality*

Neural LM

- Solves curse of dimensionality; how?
- The number of parameters are not $P(W_N|W_0W_1W_2W_3\ldots W_{N-1})$s
- The parameters are weights and biases in the neural net

Neural LM: #parameters

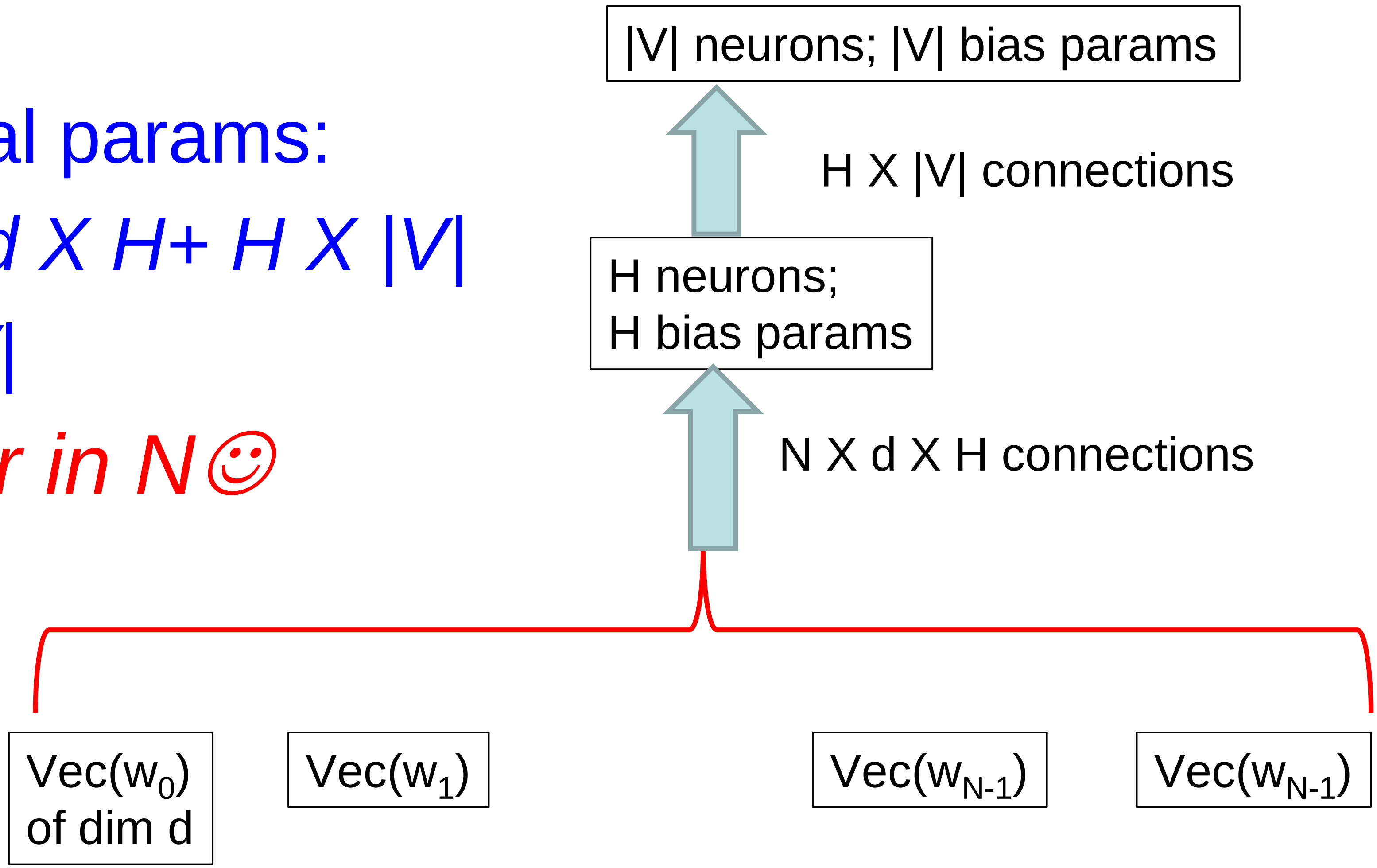
- Solves curse of dimensionality; how?

- Total params:

$=N \; X \; d \; X \; H + H \; X \; |V|$

$+H+|V|$

Linear in N☺



LLMs

- All they do is *predict next word/sentence and fill in the gaps*
- Self supervised learning
- Set the parameters for prediction

- HUGE number of parameters
  - 1 trillion (openAI) to 4 trillion (Amazon’s Olympus)
  - Aside: Olympus expenditure- $4billion, about 33K Cr INR; Goa budget about 25K Cr INR

Piggybacking on LLMs

Use the following in increasing order of resource demand for specific tasks

(a) Prompt engineering (least resource hungry)

(b) Adapter n/w

(c) Fine-tuning

(d) Pre-training

3 stages of LLM based CAI (chatGPT)

- Generative Pretraining (GP)
- Supervised Fine Tuning (SFT)
- Reinforcement Learning based on Human Feedback (RLHF)

Evolution of LLMs

- Perceptron (1960s)
- Feedforward n/w and backpropagation (1984)
- Recurrent Neural Net (RNN; 80s)
- Attention (2015)
- Transformers (2017)
- LLMs (GPTs 2020s)

But language models are very old!! 2000 years- the notion of grammar

Timeline of Language Modeling

1948

N-gram Model : predict the next word based on the previous N-1 words

1954

Distributional Hypothesis : A word is characterized by the company it keeps

1986

Recurrent Neural Networks : Processes sequential data by using the output from previous steps as inputs for the current step

2013

Word2Vec : Distributed word representation in NLP models

2014

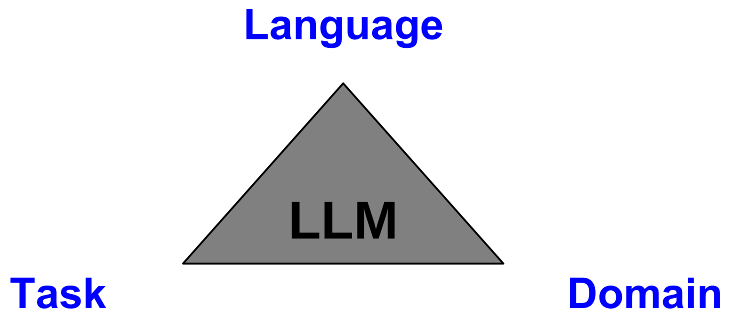
Attention : At each time step, the model selectively focuses on relevant words in the sequence

2017

Transformers : Uses attention and positional encoding to learn context-aware representations

Trinity Language Models

- General purpose LLMs need long training time and huge compute resources
- Developing specialized small language models that are easier to build and deploy
- Trinity models, e.g., *Domain: agriculture, Task: Question Answering, Language: Konkani*



Instances of Trinity Models

Legal Llama project

(In collaboration with Meta and NLSIU)

Motivation

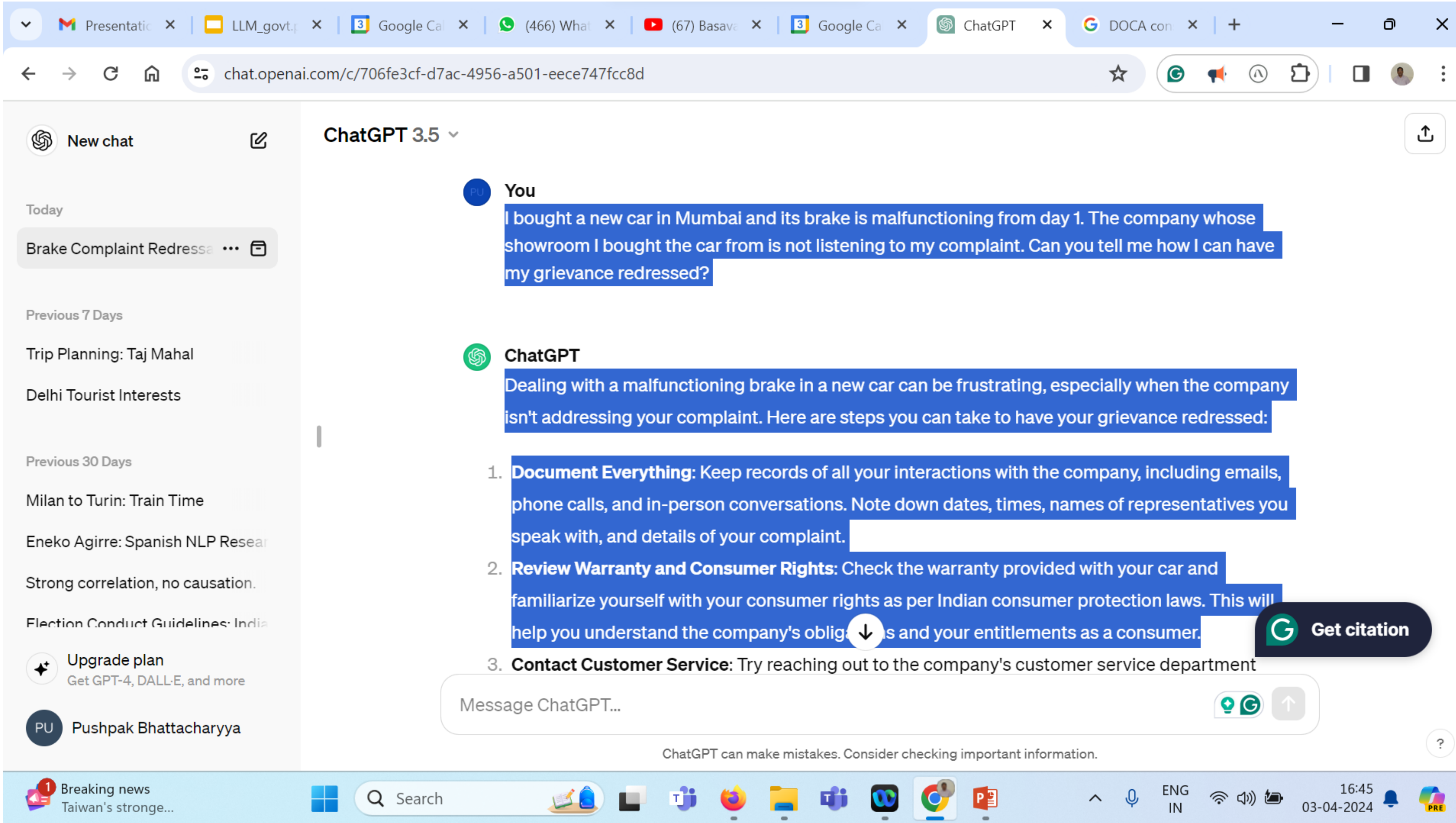
- Consumer legal affairs take up roughly 4% of the legal actions taken in India.

- **Consumer-related challenges:**
    - Lack of quick and proper guidance for grievance redressal
    - Slow addressal

- **Technology-related challenges:**
    - Current LLMs do not provide specifics
    - Often refers to incorrect laws

PU

ChatGPT



chat.openai.com/c/706fe3cf-d7ac-4956-a501-eece747fcc8d
New chat
Today
Brake Complaint Redress
Previous 7 Days
Trip Planning: Taj Mahal
Delhi Tourist Interests
Previous 30 Days
Milan to Turin: Train Time
Eneko Agirre: Spanish NLP Resea
Strong correlation, no causation.
Upgrade plan
Get GPT-4, DALL·E, and more
Pushpak Bhattacharyya
ChatGPT 3.5
You
I bought a new car in Mumbai and its brake is malfunctioning from day 1. The company whose showroom I bought the car from is not listening to my complaint. Can you tell me how I can have my grievance redressed?
ChatGPT
Dealing with a malfunctioning brake in a new car can be frustrating, especially when the company isn't addressing your complaint. Here are steps you can take to have your grievance redressed:
1. Document Everything: Keep records of all your interactions with the company, including emails, phone calls, and in-person conversations. Note down dates, times, names of representatives you speak with, and details of your complaint.
2. Review Warranty and Consumer Rights: Check the warranty provided with your car and familiarize yourself with your consumer rights as per Indian consumer protection laws. This will help you understand the company's oblig s and your entitlements as a consumer.
3. Contact Customer Service: Try reaching out to the company's customer service department
Get citation
Message ChatGPT...
ChatGPT can make mistakes. Consider checking important information.
Breaking news
Taiwan's stronge...
Search
ENG
IN
16:45
03-04-2024

Actual help…

gama.gov.in/ConsumerGrivences.aspx#:~:text=A%20dissatisfied%20consumer%20can%20file,National%20Consumer%20D...

3. **How to approach National Consumer Disputes Redressal Commission**

A dissatisfied consumer can file a complaint directly with the national commission or appeal against decisions of the state commission within a month from the date of the order. The court fee is Rs 5,000 and the demand draft should be in the name of The Registrar, National Consumer Disputes Redressal Commission. There is no fee for filing an appeal before the state or national commission. You can appeal against the orders of the national commission in the Supreme Court within a period of 30 days.

Sl. No.	Total Value of goods or services and the compensation claimed	Amount of fee payable
	District Forum	
(1)	Upto one lakh rupees – For complainants who are under the Below Poverty Line holding Antyodaya Anna Yojana Cards	Nil
(2)	Upto one lakh rupees – For complainants other than Antyodaya Anna Yojana card holders.	Rs.100
(3)	Above one lakh and upto five lakh rupees	Rs.200
(4)	Above five lakh and upto ten lakh rupees	Rs.400
(5)	Above ten lakh and upto twenty lakh rupees	Rs.500
	State Commission	
(6)	Above twenty lakh and upto fifty lakh rupees	Rs.2000
(7)	Above fifty lakh and upto one crore rupees	Rs.4000
	National Commission	
(8)	Above one crore rupees	Rs.5000

https://consumerhelpline.gov.in/about-portal.php

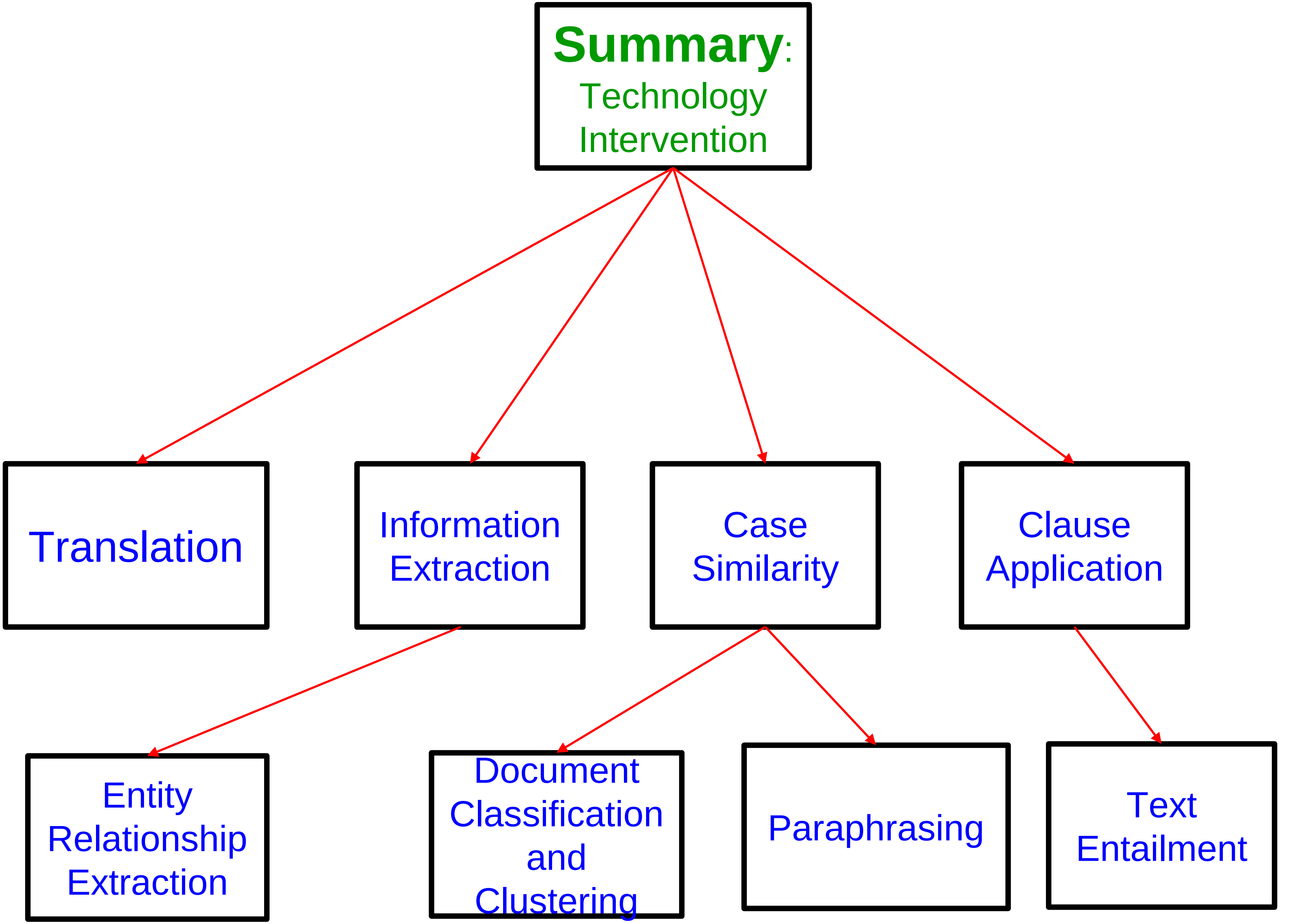
AIM of the project

- **Citizen centric Chatbot:**
  - Assist citizens in obtaining helpful responses for standard consumer queries
  - Provide guidance regarding legal procedures
- **Decision-assist tool:**
  - Help **compile information**
  - summarize complex legal data from particular grievances
  - Inferencing

Technique

- Pre-trained llama-2 language model.
- **Finetuning:-**
  - Unsupervised fine-tuning with a general legal corpus
  - Supervised fine-tuning on simulated conversations

- **Retrieval-Augmented Generation:-**
  - Train a retrieval model to obtain relevant information from the knowledge dataset for a particular query
  - Feed the information to the language model while generating the answer


Summary: Technology Intervention
Translation
Information Extraction
Case Similarity
Clause Application
Entity Relationship Extraction
Document Classification and Clustering
Paraphrasing
Text Entailment

RailLLM

In collaboration with Centre for Railway Information Systems (CRIS) and META

Problem statement

- **Smart Chatbots/Intelligent Virtual Assistants**: Answer questions about schedules, fares, platforms, and cancellation rules, and suggest tourist destinations. These services can be integrated with IRCTC website which is used by railway customers for passenger reservations.

- Providing **translation services** to different railway chatbots to ensure the availability or integration of services in regional languages for wider accessibility (such as digitally less experienced customers).

Problem statement

- **Feedback Analysis**: Collect feedback on new products and services. Whenever a product is launched, a feedback system should be incorporated into the application. Sentiment analysis or feedback categorization/clustering can be done to improve the application/product.

- **Improved Complaint Handling**:Learn from all social media, complaints, and Rail Madad data.Filter complaints to specific departments for faster resolution.Analyze sentiment and categorize feedback for better decision-making.

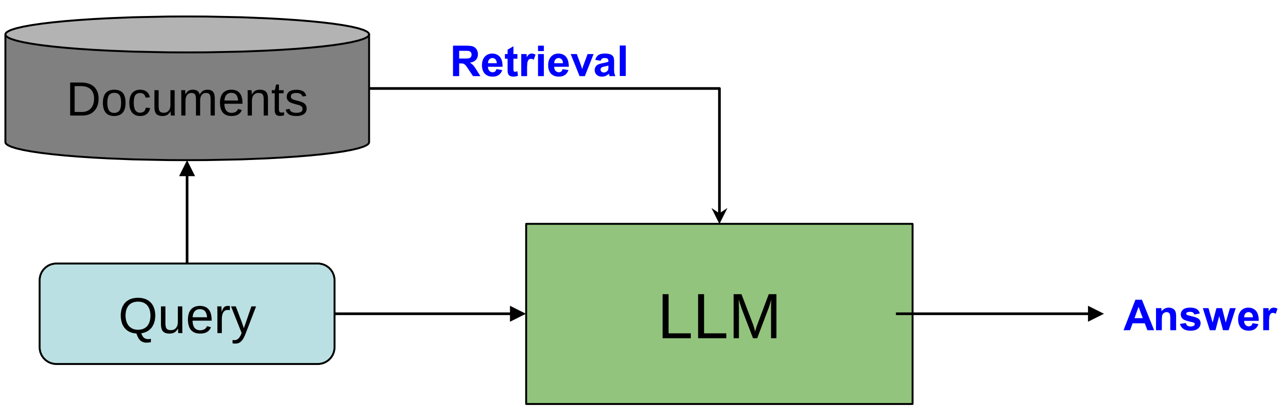
Enhancing Government Transparency and Citizen Engagement with Chat Applications

(In collaboration with Orgpedia)

Problem Statement

- Build a QA system in marathi on top of Indic language LLM for answering queries related to resolutions passed by the Maharashtra govt.

- Models used: *mT5, IndicBART, Sarvam-openhathi, Google Gemma-2B*



Small-Medium-Large Language Models

Well known LMs: size

- **GPT-3 (Generative Pre-trained Transformer 3)** by OpenAI: 175 billion parameters; **GPT-4:** 1 trillion
- **BERT** by Google: BERT-base: 110 million; BERT-large: 340 million
- **XLNet** by Google AI and Carnegie Mellon University: Size: 340 million
- **T5 (Text-To-Text Transfer Transformer)** by Google AI: Size: 11 billion
- **LLAMA-2** by Meta**:** 34 billion
- **Olympus** by Amazon: 4 trillion
- **Mistral 7B** by Mistral.AI: 7.3 billion
- **OpenHathi** by Sarvam: 7 Billion
- **PaLM** by Google: 540 billion

Comparison (1/2)

- **#Parameters**
  - The human brain has some 8.6 x **10^{10} (eighty six billion) neurons**. Each neuron has on average 7,000 synaptic connections to other neurons.
  - Hence total no. of “parameters”= average **600 trillion**

Comparison (2/2)

- **Energy Requirement**
  - Human: approximately 175.2 KWh/year
  - A typical household in India: 260 KWh/mo=3120 KWh/year
  - GPT3: 1647 MWh/year=1647000 KWh/year (about 10000 times that of a human and about 500 times that of an Indian household)
- **Carbon Footprint** (in CO_2e)
  - Human: global average of 4500 KG/year/person
  - GPT3: 430 to 680 metric tons, i.e., 430000 to 680000 KG/year (about 100 times)

Neurophysiology: Broca’s Area

Broca’s area in frontal lobe, left hemisphere

Damage messes up syntax

a-grammatical speech production; Inability to use syntactic information; telegraphic speech; loss of function words and suffixes

“I eat rice spoon”

→ *I eat rice with spoon*

No difference between

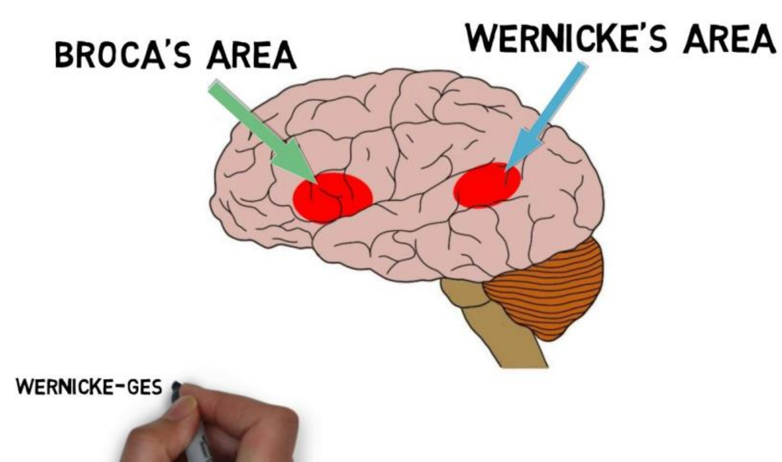
- *Visit to the President vs.*
- *Visit by the President*

Wernicke's Area

Wernick's area is in superior temporal gyrus, left hemisphere

Damage messes up semantics-pragmatics

Fluent meaningless phrases: *The pink elephant sang and the blue stone danced*



A few SLMs

Credit: https://analyticsindiamag.com/9-best-small-language-models-released-in-2023/

A few “Small” Language Models (SLMs): 1/3

- **Llama-2 7B:** many current open source models built on top of Llama family of models; multilingual; various NLP tasks
- **Phi2 and Orca**: 13 B; Microsoft; Reasoning and explainability capability
- **Stable Beluga 7B**: leverages Meta’s Llama; multilingual: text generation, translation, question answering, and code completion.

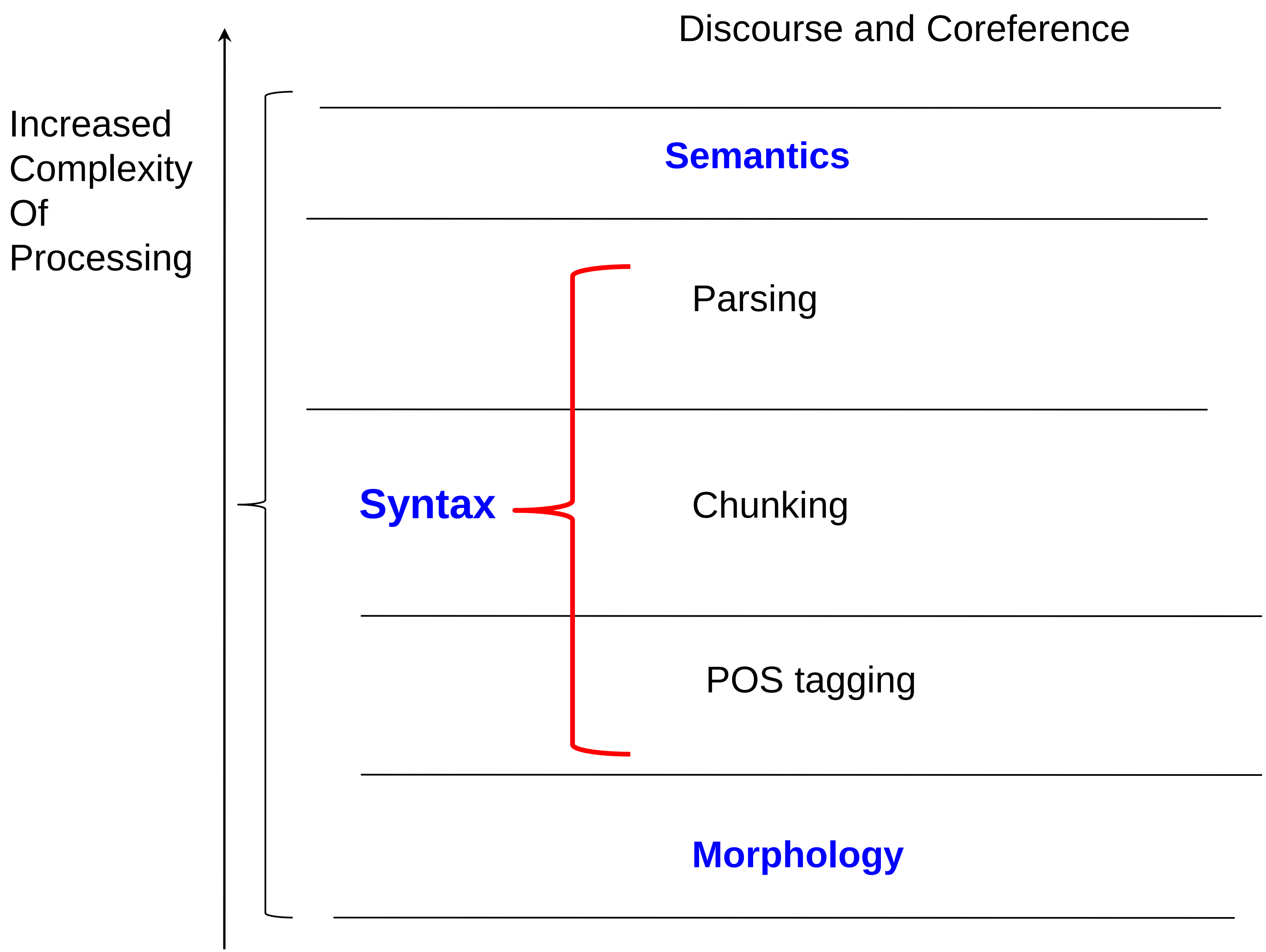
SLMs: 2/3

- **X Gen**: 7B; by Salesforce; multilingual; creative writing and content creation and language learning
- **Alibaba's Qwen**: Qwen-1.8B, Qwen-7B, Qwen-14B, and Qwen-72B; standard NLP tasks, and audio processing
- **Alpaca 7B**: leverages Meta; renowned for its remarkable compactness and cost-effectiveness, requiring less than $600 in building costs

SLMs: 3/3

- **MPT**: 7 B; by Mosaic ML; creative writing, content creation, education, and accessibility tools
- **Falcon 7B**: by Technology Innovation Institute (TII); Tailored for chatting and question answering; tops the Huggingface leaderboard for the longest time
- **Zephyr**: 7B; by Huggingface; specialized for chatbots

NLP Layers


Discourse and Coreference
Increased
Complexity
Of
Processing
Semantics
Parsing
Syntax
Chunking
POS tagging
Morphology

Our NLP Lab at IIT Bombay*: Since 2000;* Works in all areas of NLP- **translation, QA, Sentiment and Emotion, Natural Language Generation** and so on.

Key Research Areas

Machine Translation

Sentiment Analysis

Information Retrieval

Lexical Semantics

Information Extraction

Cognitive NLP