Foundations of NLP

Introduction
CS3126

Course Instructor : Nidhi Goyal

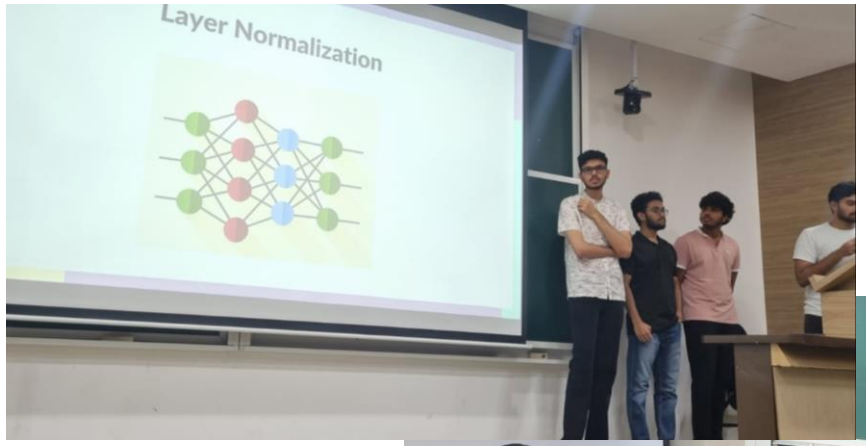
Mahindra™ University
Global Thinkers. Engaged Leaders.

Who am I?

- **Research interests**
  - NLP, Knowledge Graphs, Graph Neural Networks, Computational Neuroscience, Computer Vision, Brain Computer Interfaces, Generative AI and solving real-world problems
- **Experience (Research and Teaching)**
  - Applied researcher in NLP, Knowledge Graphs, Graph Neural Networks ~ 6 years [Research scholar -IIIT-Delhi and Naukri.com]
  - Taught NLP last semester (CS3216/AI5203) at Mahindra University
  - Former Assistant Professor, GGSIPU, New Delhi
  - Former Assistant Professor, DCE, Gurgaon
  - TA for many courses- COA, Privacy and security on Online social media @IIIT, NPTEL, NLP course PGDSAAI@IIIT-Delhi

Nidhi Goyal
(Assistant Professor, Ecole School of Engineering)

Previous Course Offerings (NLP) /CS3216/AI5203



Lecture Plan

- Course Logistics
- Course Outcomes
- Content/Syllabus
- Grading Policy
- Why should you learn NLP?
- What is Natural Language Processing?
- Applications
- Resources

Course Logistics

- Attendance is mandatory!!
- All attendance is via QR codes
- *"Missing Minor exams will not be given re-test. If a student misses it due to genuine and verifiable reason(s), she/he will be given weightage for the missed exam(s) on a pro-rate basis of performance in the end-sem exam."*

- Lecture Timings: Mondays, Wednesdays & Thursdays- 10:35–11:30 AM
  (Strictly adhere to class timings)
- Slack, help sessions/office hours (for all course questions/discussion)
  - Slack for assignments, doubts, course announcements, deadlines, etc.
  - Office hours start **Monday and Wednesday (4:00-5:00 pm)**, Location- **IT2 block, 180-seater lab, my cabin**
- For any queries: *nidhi.goyal@mahindrauniversity.edu.in*
  - Use email judiciously for important/urgent communication
- Follow Course page for resources, materials, etc. We've put a lot of other important information on the class webpage (Link to be published soon).
- Slide PDFs and other resources shall be uploaded before each lecture.

Smooth functioning of Course

- Slack for assignments, doubts, course announcements, deadlines, etc.
  - Create your account on slack with Your Roll number as username.
  - Get yourself added to below channels by your instructor or TAs
    - nlp-announcements
  - Be active to check slack notifications for course related information!!
  - First communication point for doubts on slack (#doubts channel).
  - Healthy discussions among peers lead to effective learning.

Course Outcomes

After the successful completion of this course , the students will be able to:

- CO1: Understand the fundamental concepts in Natural Language Processing, algorithms, methods, and strategies.
- CO2: Analyze and evaluate the strengths and weaknesses of various NLP techniques.
- CO3: Apply NLP tools and libraries, gaining hands-on experience with existing natural language processing and machine learning frameworks.
- CO4: Implement and adapt cutting-edge frameworks in NLP for practical applications.

Course Content/Syllabus

- Module-1: Introduction
- Module-2: Syntax, Regular Expressions & Text Normalization
- Module-3: Text Pre-processing and Sequence labeling
- Module-4: Probabilistic Language Modeling
- Module-5: Semantics & Word embeddings
- Module-6: Neural & RNN-based Language Modeling
- Module-7: Advanced Topics/Large Language Models and NLP Applications

Details are found here: Syllabus-NLP

Grading Policy

Type of Evaluation	% Contribution in Grade
Minor 1 and Minor 2	30 (15+15)
Majors	25
In class activity/quiz	10
Assignments	10
Project	25

Cheating Policy

https://www.mahindrauniversity.edu.in/sites/default/files/2022/Annexure-XVI-Unfair-Means.pdf

- Exams and quizzes are to be completed individually. Verbal collaboration on homework assignments is acceptable, but
  - you must not share any code or other written material,
  - Everything you submit must be your own work, and
  - you must note the names of anyone you collaborated with on each problem (the only exceptions are the instructors and TAs), and the nature of the collaboration (e.g., “X helped me,” “I helped X,” “X and I worked it out together.”).
- If you find material in published literature (e.g., on the Web) that is helpful in solving a problem, you must cite it and explain the answer in your own words.
- The project is to be completed by a team; you are not permitted to discuss any aspect of your project with anyone other than your team members, the instructor, and the TAs.
- You are encouraged to use existing NLP components in your project; you must acknowledge these appropriately in the documentation.

Lecture Ethics

What do you see here?

Diversity of languages in this world

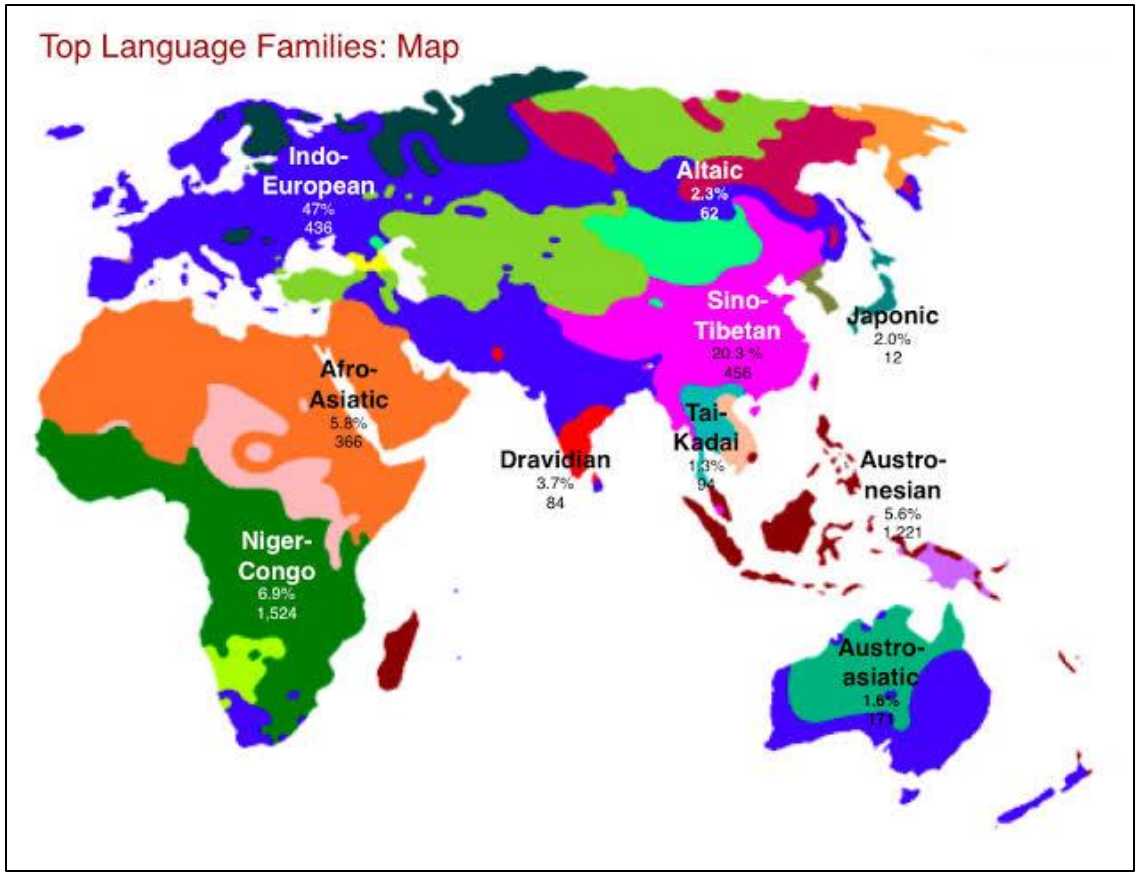
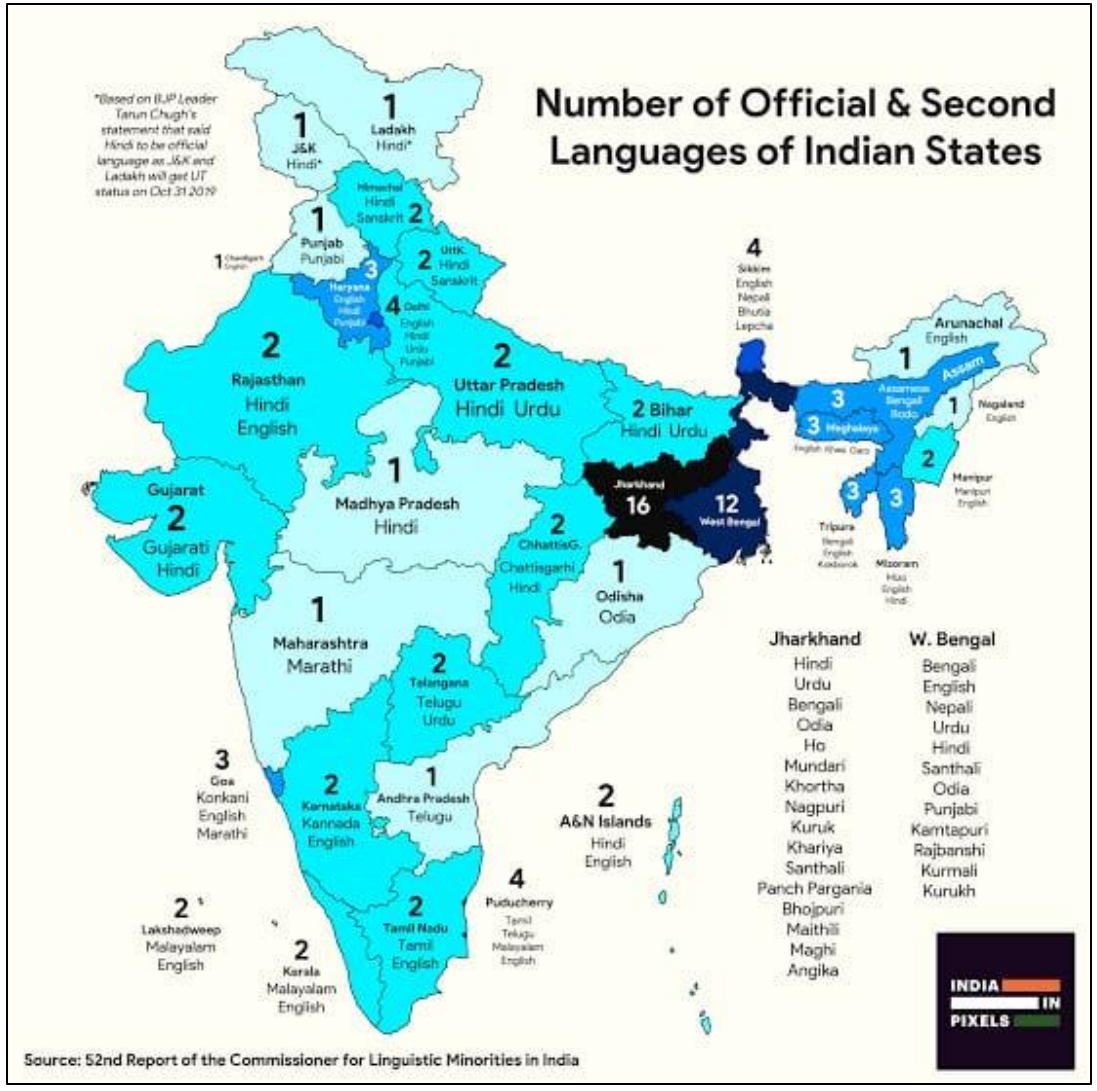


Image source: https://triangulations.wordpress.com/2013/11/22/language-families/

Variations across Indian languages



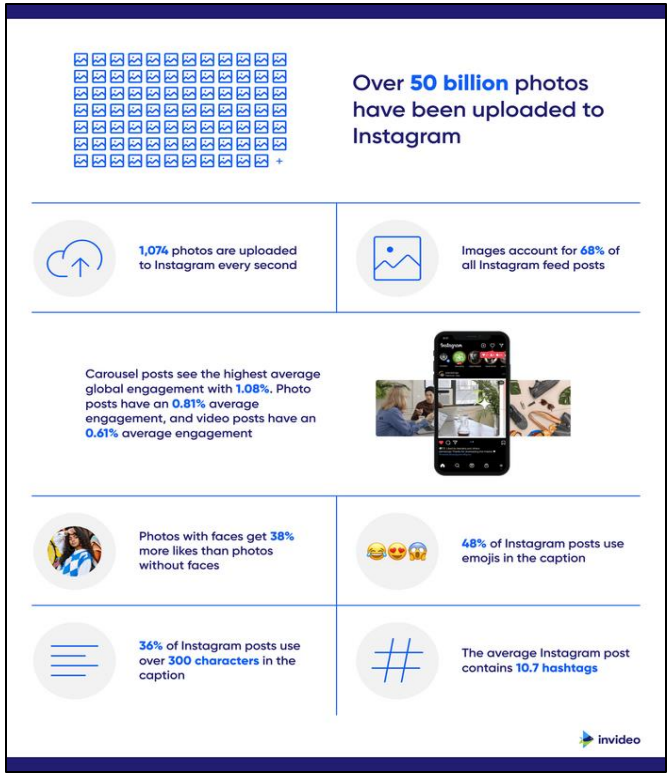
Source: https://twitter.com/indiainpixels/status/1299284220687167488

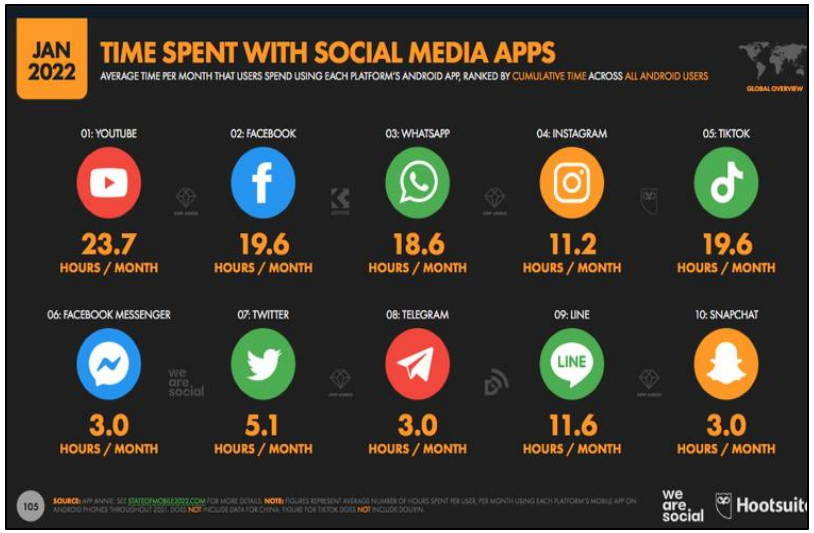
Natural language

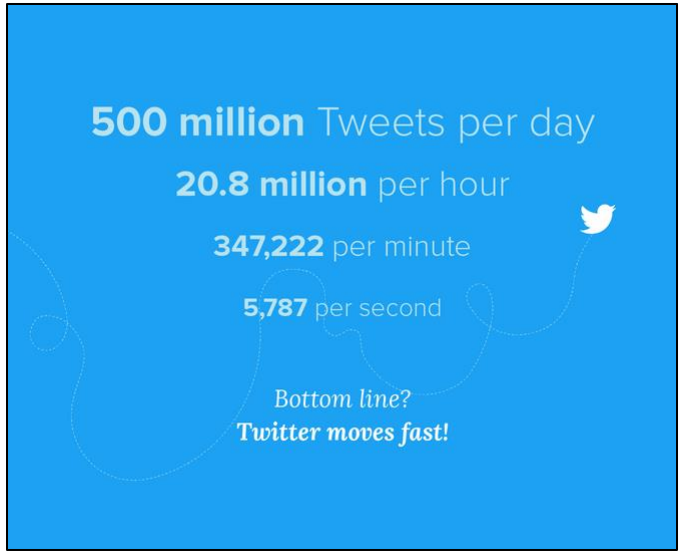
In neuropsychology, linguistics, and philosophy of language,

a **natural language** is any **language that occurs naturally in a human community by a process of use, repetition, and change without conscious planning or premeditation.**

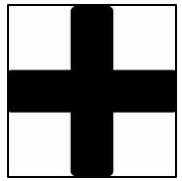
Data Data everywhere!!!!

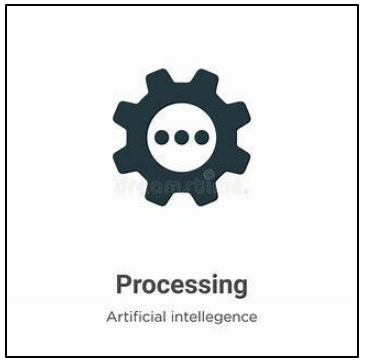






Natural language processing?





Natural language understanding

Natural language generation

Natural language processing

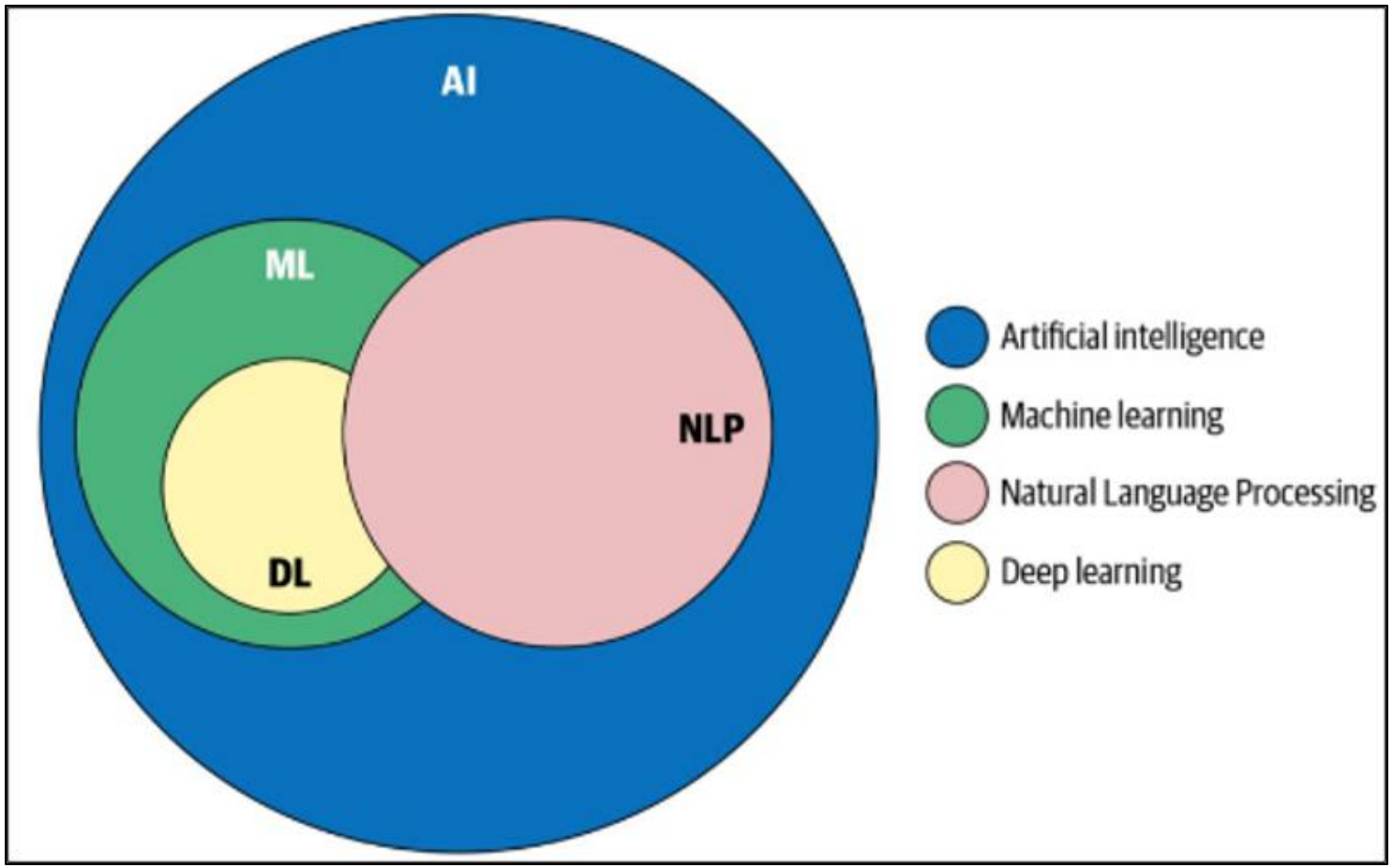
From Wikipedia, the free encyclopedia

For other uses, see NLP.
This article is about natural language processing done by computers. For the natural language processing done by the human brain, see Language processing in the brain.

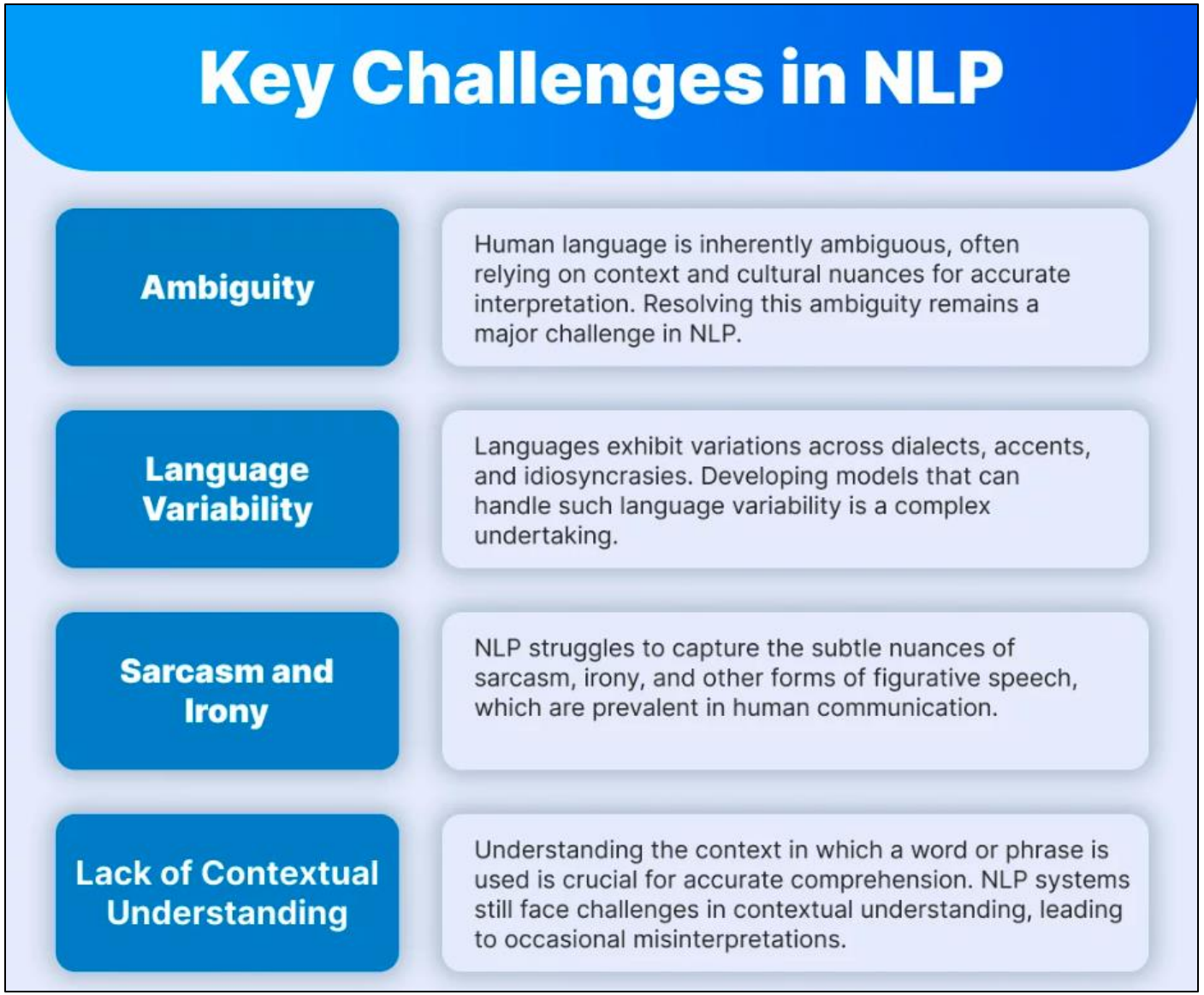
Natural language processing (**NLP**) is an interdisciplinary subfield of computer science and linguistics. It is primarily concerned with giving computers the ability to support and manipulate human language. It involves processing natural language datasets, such as text corpora or speech corpora, using either rule-based or probabilistic (i.e. statistical and, most recently, neural network-based) machine learning approaches. The goal is a computer capable of "understanding" the contents of documents, including the contextual nuances of the language within them. The technology can then accurately extract information and insights contained in the documents as well as categorize and organize the documents themselves.

Challenges in natural language processing frequently involve speech recognition, natural-language understanding, and natural-language generation.

https://en.wikipedia.org/wiki/Natural_language_processing



https://www.oreilly.com/library/view/practical-natural-language/9781492054047/


Key Challenges in NLP
Ambiguity
Human language is inherently ambiguous, often relying on context and cultural nuances for accurate interpretation. Resolving this ambiguity remains a major challenge in NLP.
Language Variability
Languages exhibit variations across dialects, accents, and idiosyncrasies. Developing models that can handle such language variability is a complex undertaking.
Sarcasm and Irony
NLP struggles to capture the subtle nuances of sarcasm, irony, and other forms of figurative speech, which are prevalent in human communication.
Lack of Contextual Understanding
Understanding the context in which a word or phrase is used is crucial for accurate comprehension. NLP systems still face challenges in contextual understanding, leading to occasional misinterpretations.

Ambiguity in NLP

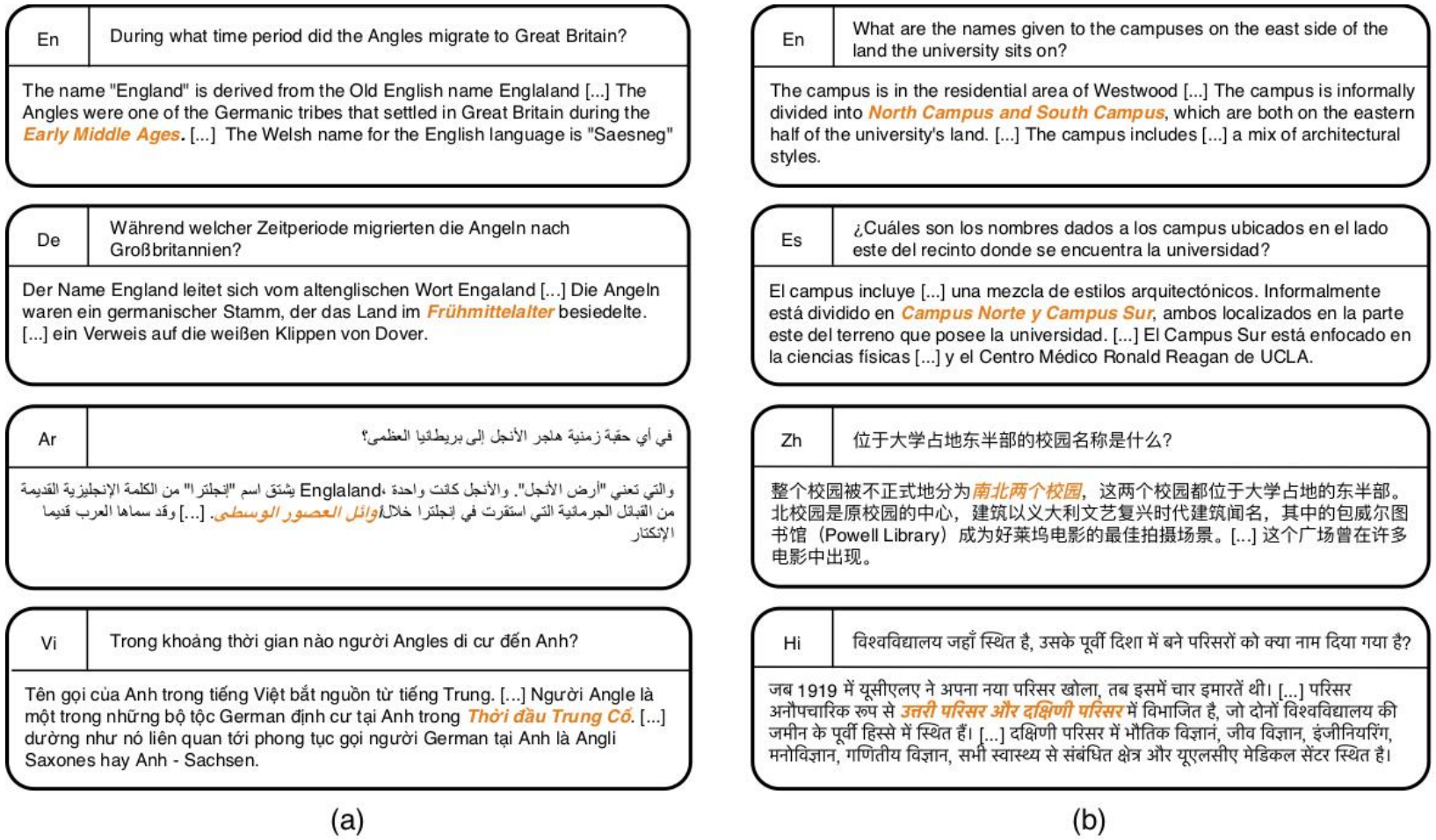
Lexical Ambiguity: This type of ambiguity represents words that can have multiple assertions. For instance, in English, the word “back” can be a noun (back stage), an adjective (back door) or an adverb (back away).

Syntactic Ambiguity: This type of ambiguity represents sentences that can be parsed in multiple syntactical forms. Take the following sentence: “ I heard his cell phone ring in my office”. The propositional phrase “in my office” can be parsed in a way that modifies the noun or on another way that modifies the verb.

Semantic Ambiguity: This type of ambiguity is typically related to the interpretation of sentence. For instance, the previous sentence used in the previous point can be interpreted as if I was physically present in the office or as if the cell phone was in the office.

Metonymy: Arguably, the most difficult type of ambiguity, metonymy deals with phrases in which the literal meaning is different from the figurative assertion. For instance, when we say “Samsung us screaming for new management”, we don’t really mean that the company is literally screaming (although you never know with Samsung these days ;)).

Multilingual



Code-mixed data

Example I

CODE-MIXED SENTENCE: is seat me girne ka koi chance nhi hai
ENGLISH TRANSLATION: there is no chance of falling down from this seat
REQUIRE CHANGES IN THE ENGLISH TRANSLATION?: No

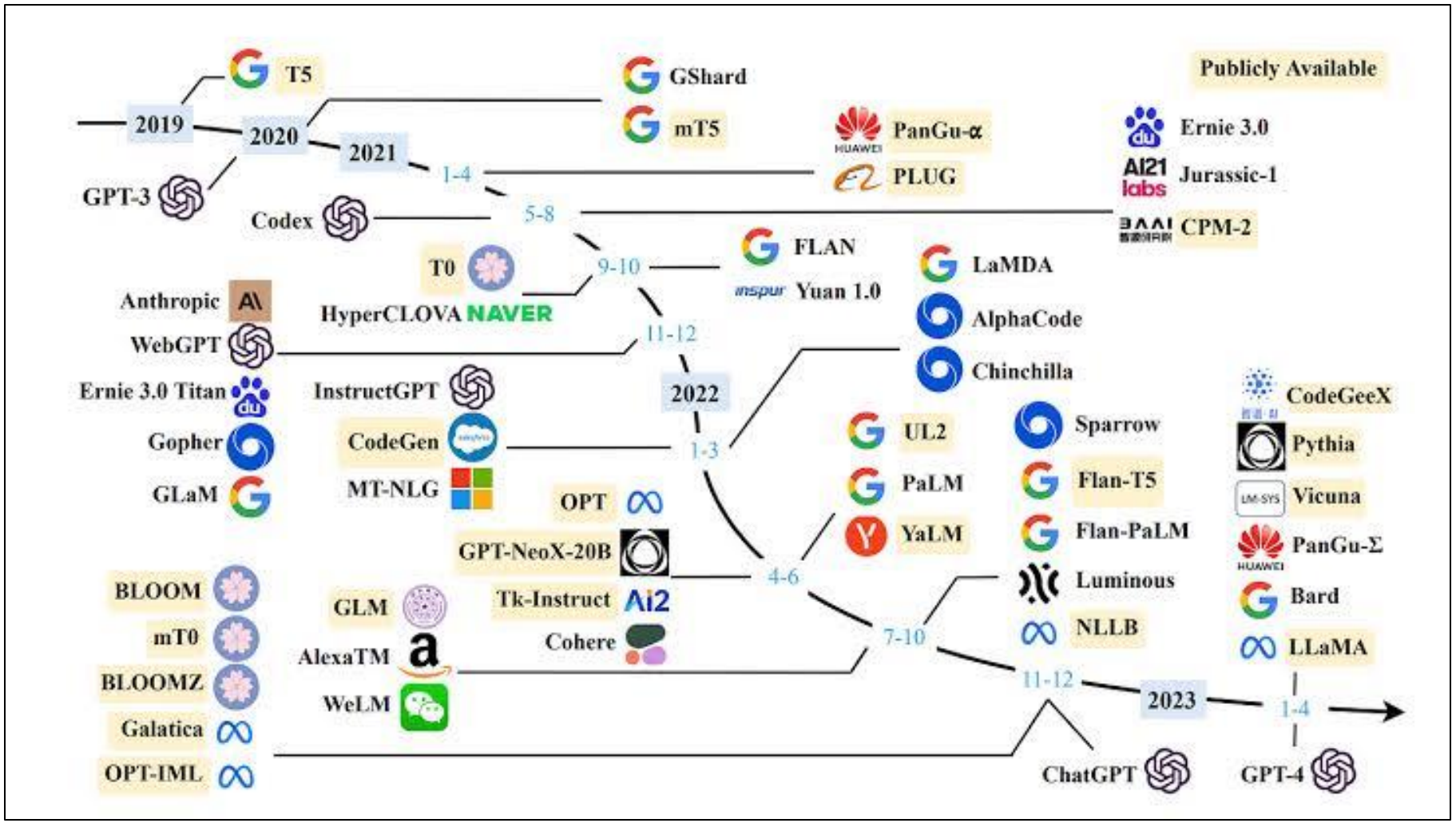
Example II

CODE-MIXED SENTENCE: Thnks buds! Kabhi kabhi aajate hai achhe photos
ENGLISH TRANSLATION: Thank you buddy, sometime good photos are captured.
REQUIRE CHANGES IN THE ENGLISH TRANSLATION?: No

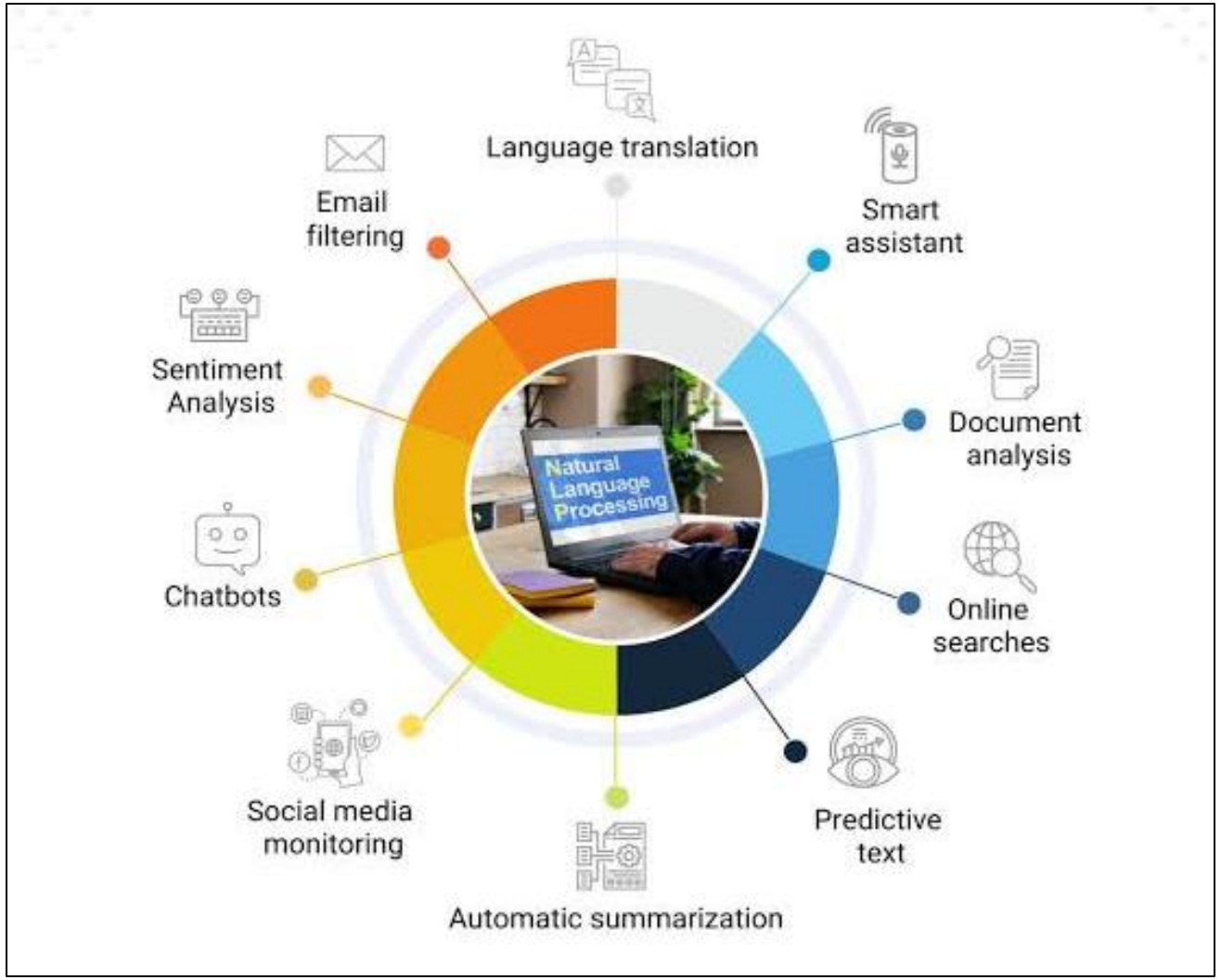
Example III

CODE-MIXED SENTENCE: Australia ke saath abhi jeete nahi hai, magar NZ ke saath final kaise jeetenge iss soch mein bhartiya yuvak on twitter.
ENGLISH TRANSLATION: Indian youth on twitter thinking that - We have not won against Australia yet, but how would we win final with NZ?
REQUIRE CHANGES IN THE ENGLISH TRANSLATION?: Yes

NLP research and advanced topics

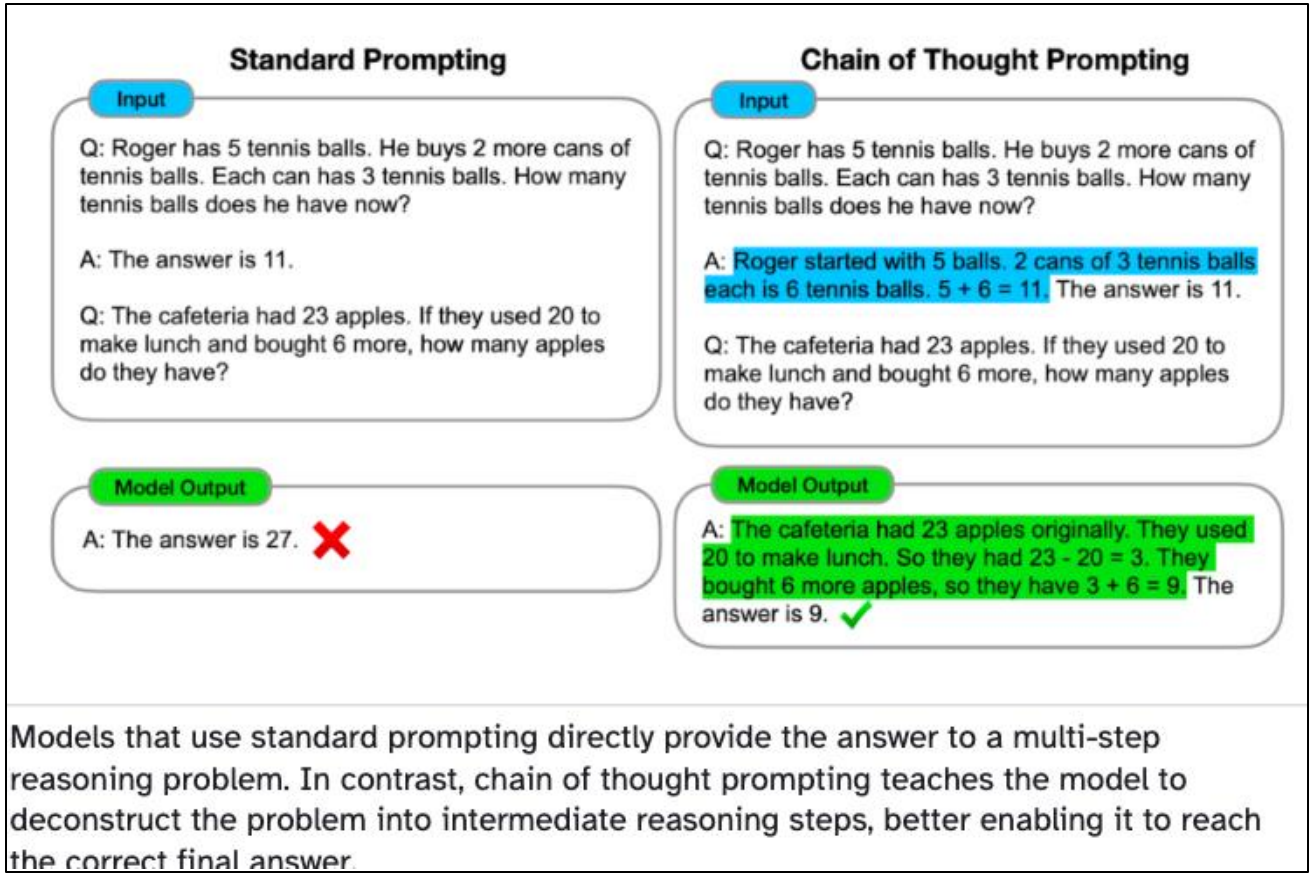


NLP Applications



https://www.youtube.com/watch?v=lj0bFX9HXeE&t=3s

Applications!



Models that use standard prompting directly provide the answer to a multi-step reasoning problem. In contrast, chain of thought prompting teaches the model to deconstruct the problem into intermediate reasoning steps, better enabling it to reach the correct final answer.

https://blog.research.google/2023/01/google-research-2022-beyond-language.html

References

https://research.google/research-areas/

https://research.ibm.com/topics/natural-language-processing

https://openai.com/research/better-language-models

Resources

- (A) Speech and Language Processing by Daniel Jurafsky and James H. Martin (3rd edition)
- (B) Natural Language Processing with Python. (updated edition based on Python 3 and NLTK 3) Steven Bird et al. O'Reilly Media

https://vlanc-lab.github.io/mu-nlp-course/teachings/fall2024-books.html