A Gentle Introduction to Generative AI (including NLP)

Centre for Artificial Intelligence (CAI)

Faculty of Engineering and Information Sciences (EIS)

UNIVERSITY
OF WOLLONGONG
AUSTRALIA

Dr Jack Yang
Web: https://uow.info/cai
Email: cai-admin@uow.edu.au

Centre for Artificial Intelligence (CAI)

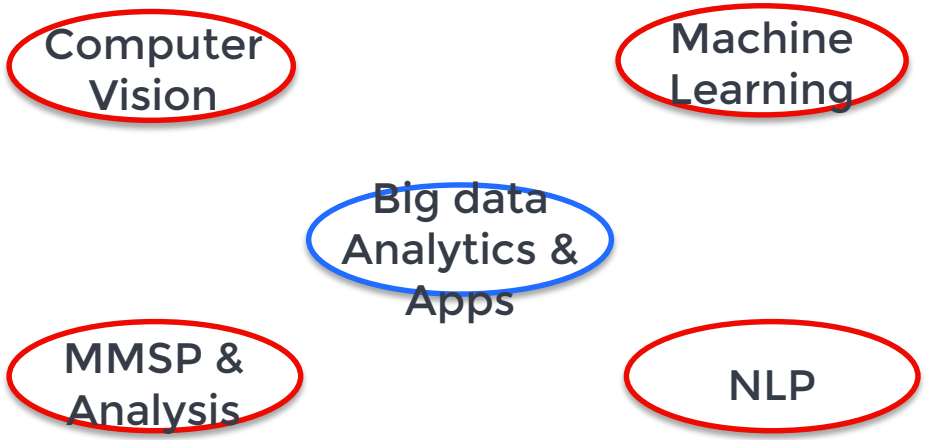
Faculty of Engineering and Information Sciences

Staff

- 3 Professors, 1 Associate Professor, 1 MCR and 1 ECR
- 1 Honorary Professor
- 3 Postdocs & RA

Research Students

- 25 (PhD, MPhil and MSc)



Outline

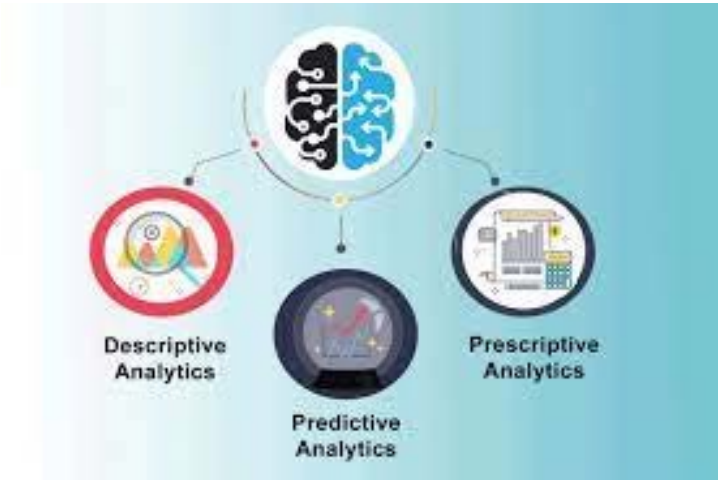
- Generative AI (GAI)
- Language Modelling
- Prompt engineering

Generative Artificial Intelligence (GAI)

Definition

Generative AI is a type of artificial intelligence technology that broadly describes machine learning systems capable of **generating text, images, code or other types of content**, often in response to a **prompt** entered by a user.

Predictive/Descriptive AI vs Generative AI (GAI)



Predictive/Descriptive AI:

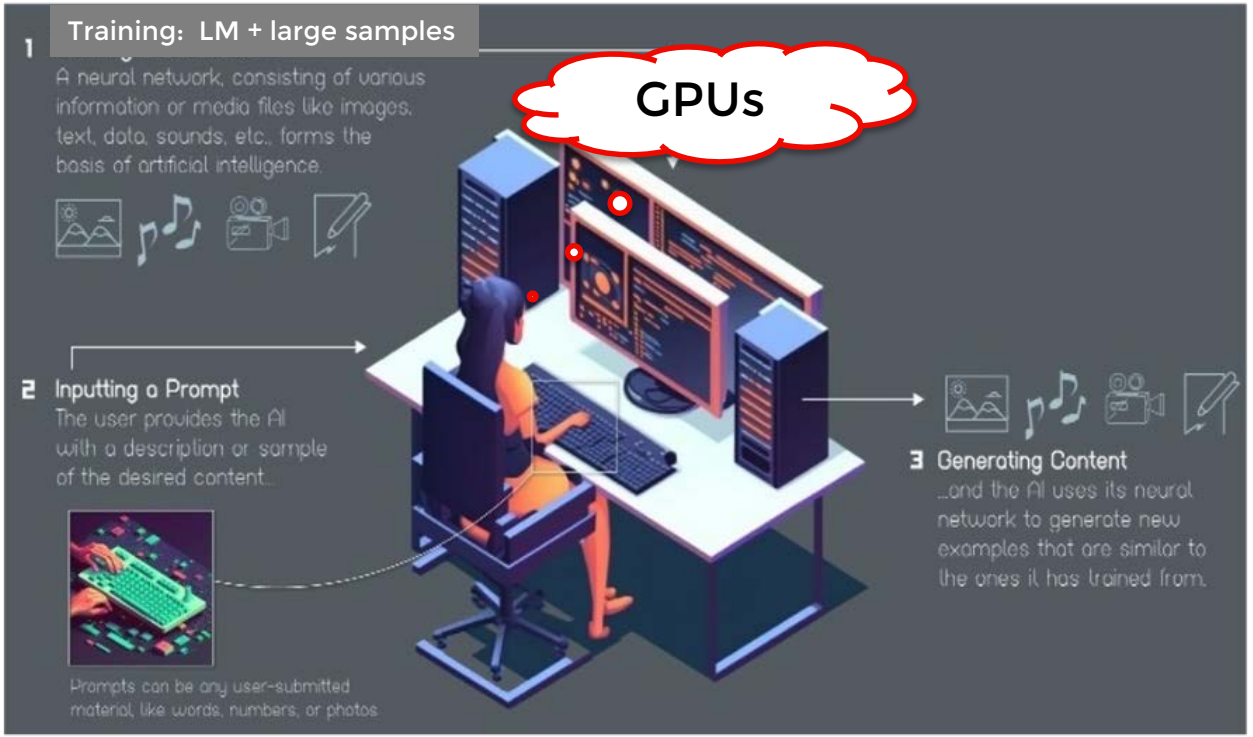
- Predict future data and identify patterns

Predictive/Generative AI

- Generate content or data

How GAI Works ?

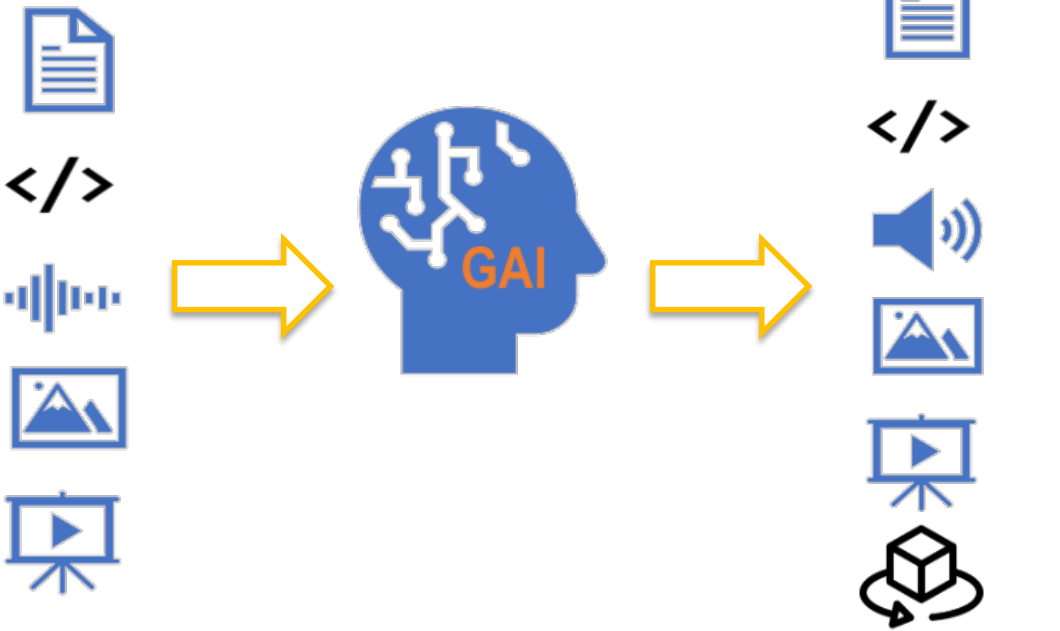
Training, prompt and inference (generation)



https://www.visualcapitalist.com/generative-ai-explained-by-ai/

How GAI Works?

- **From prompt to output**



Applications

- Personal Assistant
- Auditor
- Advisor
- Designer
- Software Engineer
- Game Creators
- Artist
- Painter
- Tutor
- Writer
- "Live" Encyclopedia
- ...

Image Generation and Editing

DALL-E, Midjourney & Stable Diffusion

Prompt: Vibrant California poppies

DALL-E

Midjourney

DALL-E

Image editing

Fake Images




Music/Audio Generation

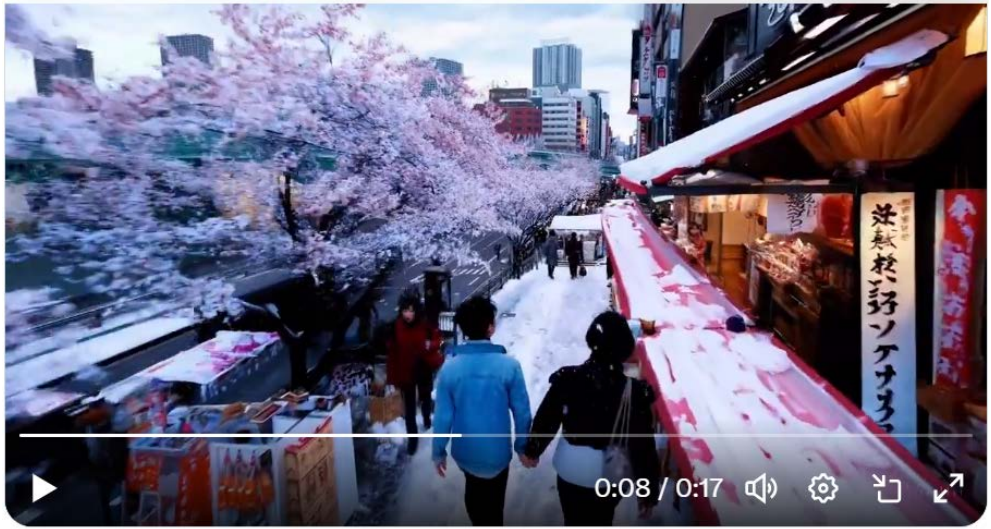
Google's MusicLM

"Meditative song, calming and soothing, with flutes and guitars. The music is slow, with a focus on creating a sense of peace and tranquility."

"jazz with saxophone"

Sora

Create realistic and imaginative scenes (Video)



https://twitter.com/i/status/1758192957386342435

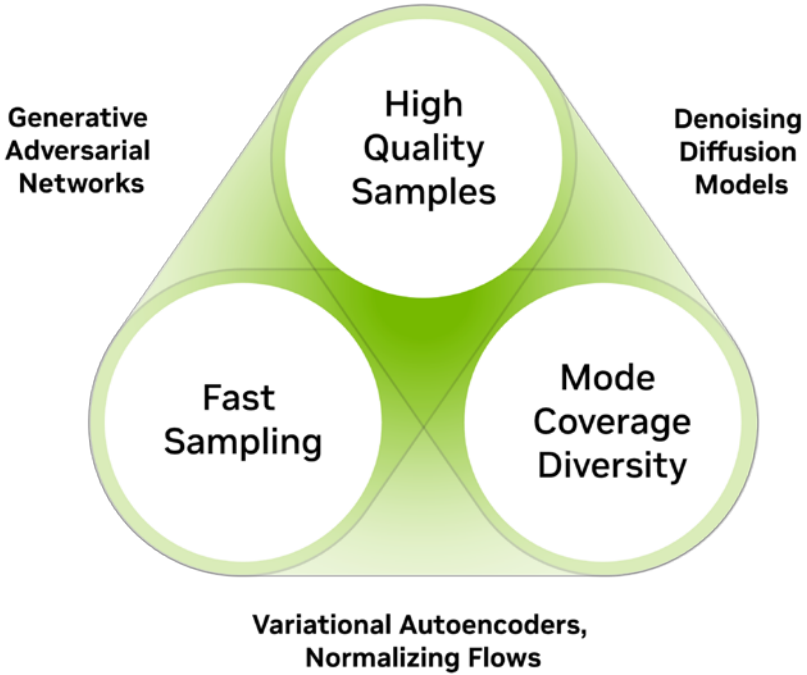
Text Generation

ChatGPT

- **Answering Questions**: I can provide answers on a broad range of topics, from science and history to technology and pop culture.
- **Tutoring:** I can help explain complex concepts in subjects like mathematics, physics, literature, and more.
- **Programming Help:** I can help debug code, provide code examples, and explain programming concepts.
- **Recommendations:** I can suggest books, movies, music, etc., based on your preferences.
- **Writing Assistance:** I can help you draft, edit, or proofread essays, stories, emails, and more.
- **Translation:** I can translate sentences or short paragraphs across various languages.
- **Conversations:** I can engage in general chat or help talk through complex ideas.
- **Tasks Involving Logic**: I can solve puzzles, math problems, and logic games.
- **Image Analysis:** With my new image input capabilities, I can analyze and describe images, though my capabilities in this regard are basic.
- **Guidance on Various Topics:** From cooking recipes to DIY instructions, fitness advice, or travel suggestions.
- **Simulation of Characters:** I can simulate dialogues from fictional or historical characters based on available knowledge.
- **Meditation and Relaxation:** I can guide you through breathing exercises or provide calming narrations.

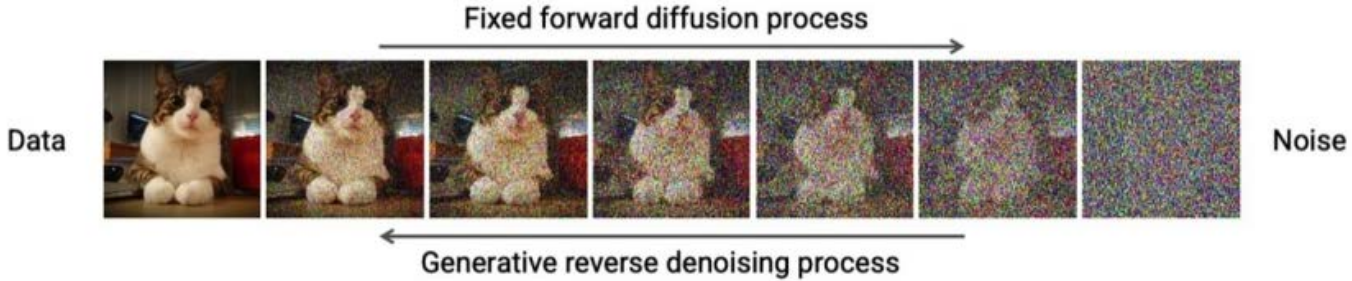
How GAI Works?

Three requirements and common architectures



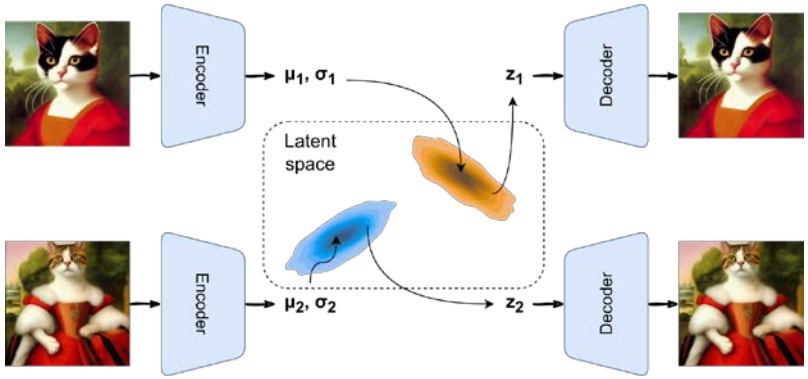
How AGI Works

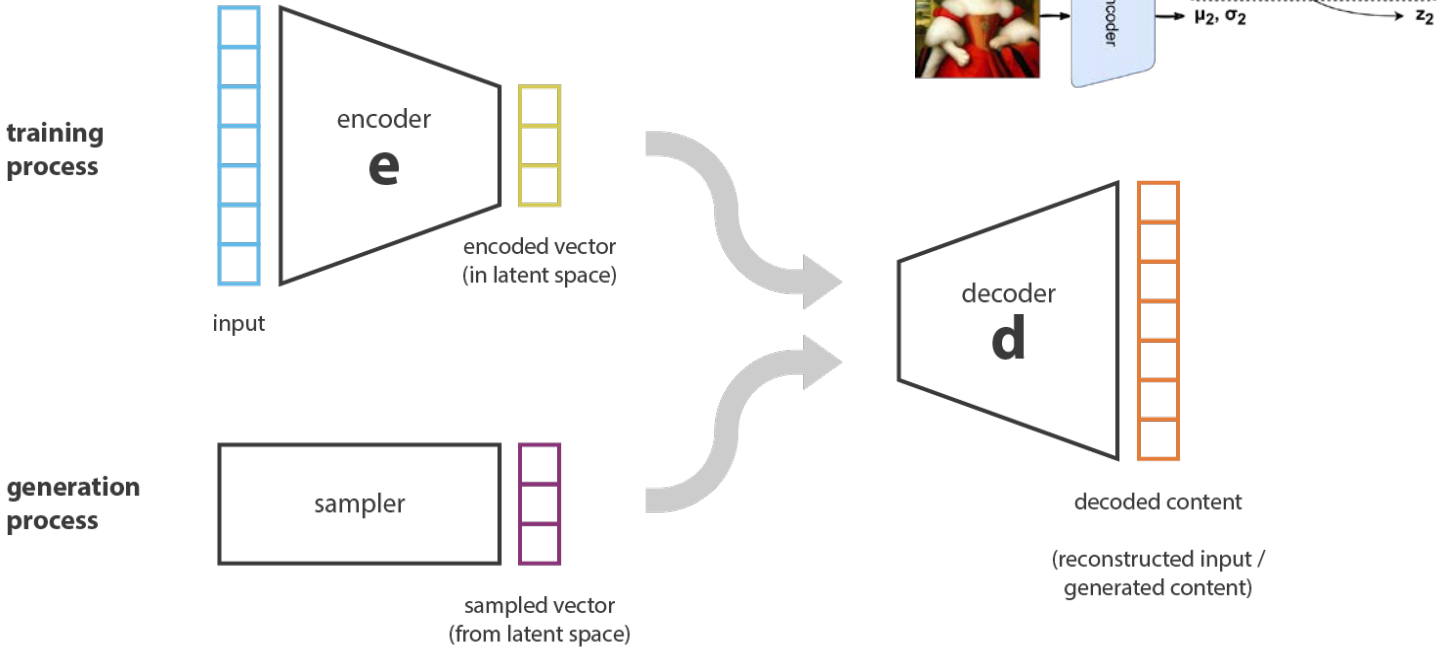
Diffusion Models



How AGI Works

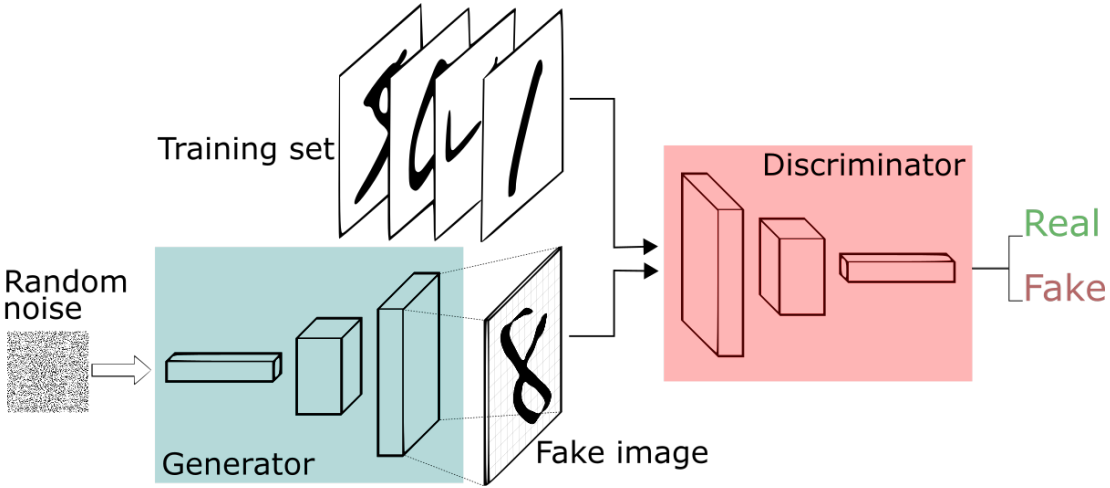
Variational autoencoders (VAEs)

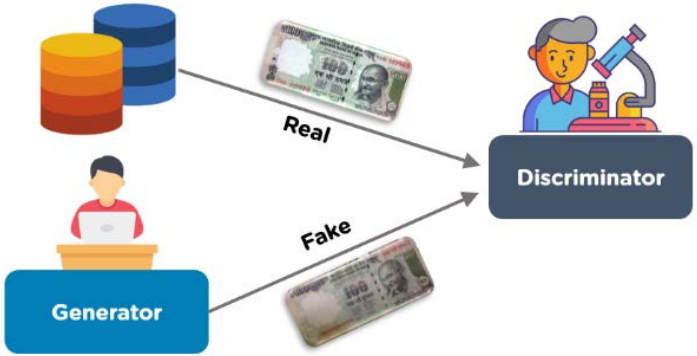




How GAI Works?

Generative Adversarial Networks (GANs)

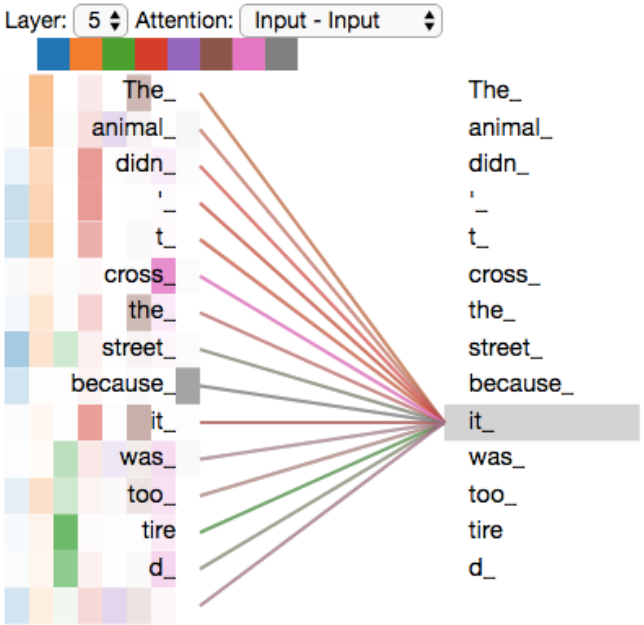




How GAI Works

Transformer - Most common encoder/decoder architecture

Two mechanisms : self-attention and positional encodings (more later)



Language Modelling

UNIVERSITY
OF WOLLONGONG
AUSTRALIA

ChatGPT – GAI

A milestone in NLP !

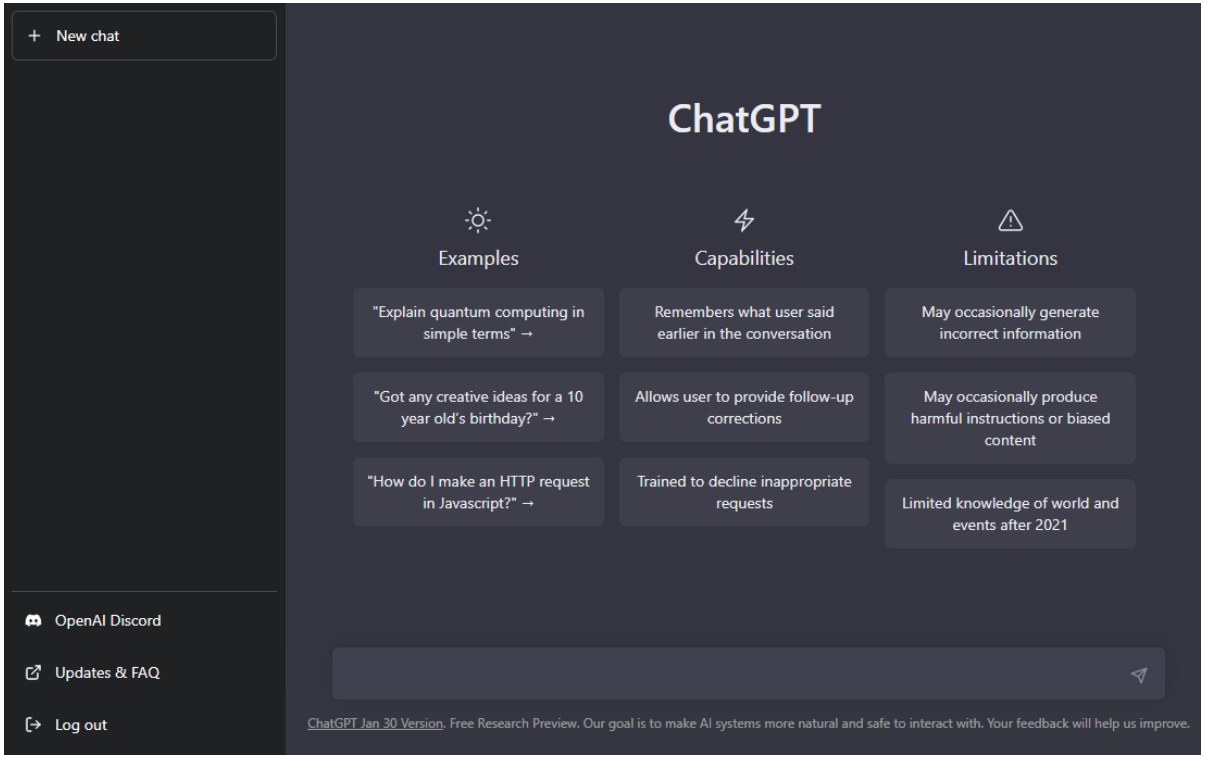
Released

- 30 Nov 2022

One million users

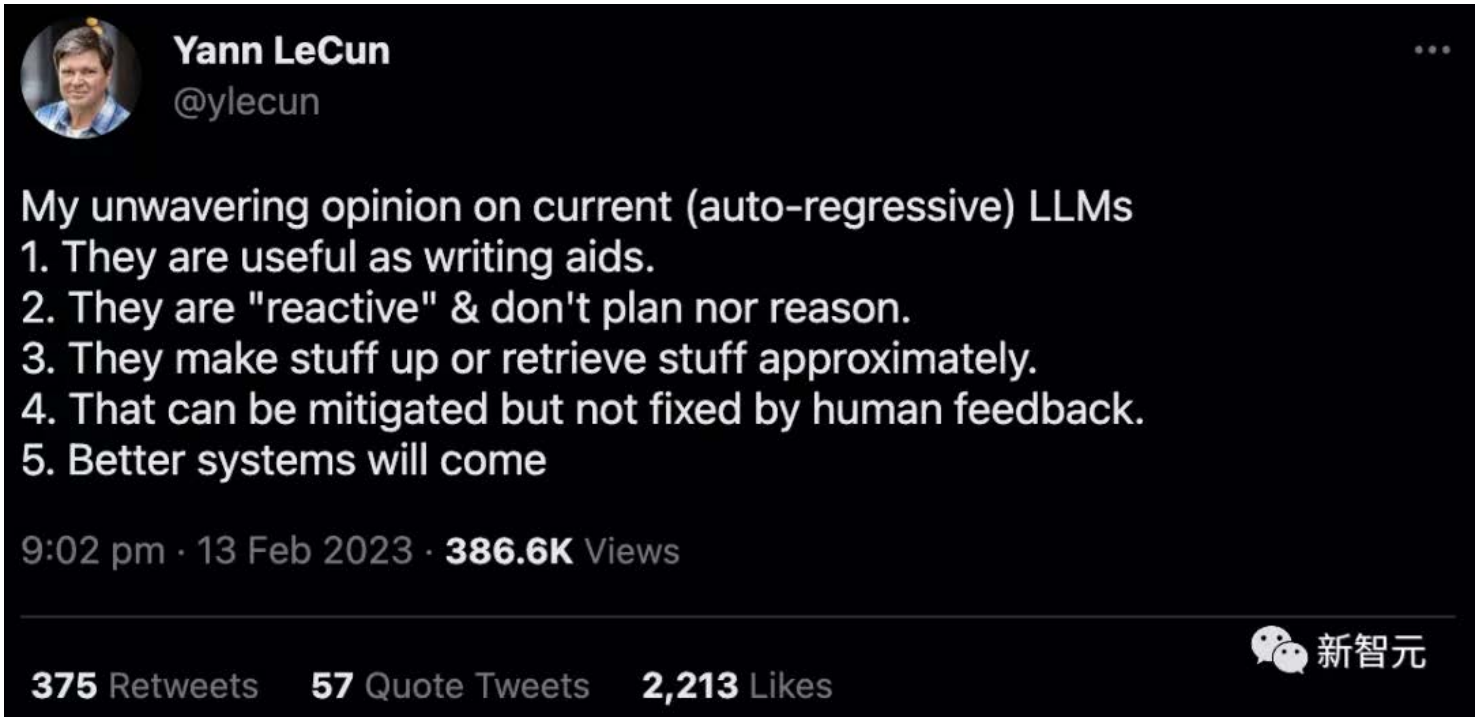
- ChatGPT 5 days
- Instagram 2.5 months
- Spotify 5 months
- Facebook 10 months
- Netflix 3.5 years

Roughly 1.5 billion people are using chatbots,

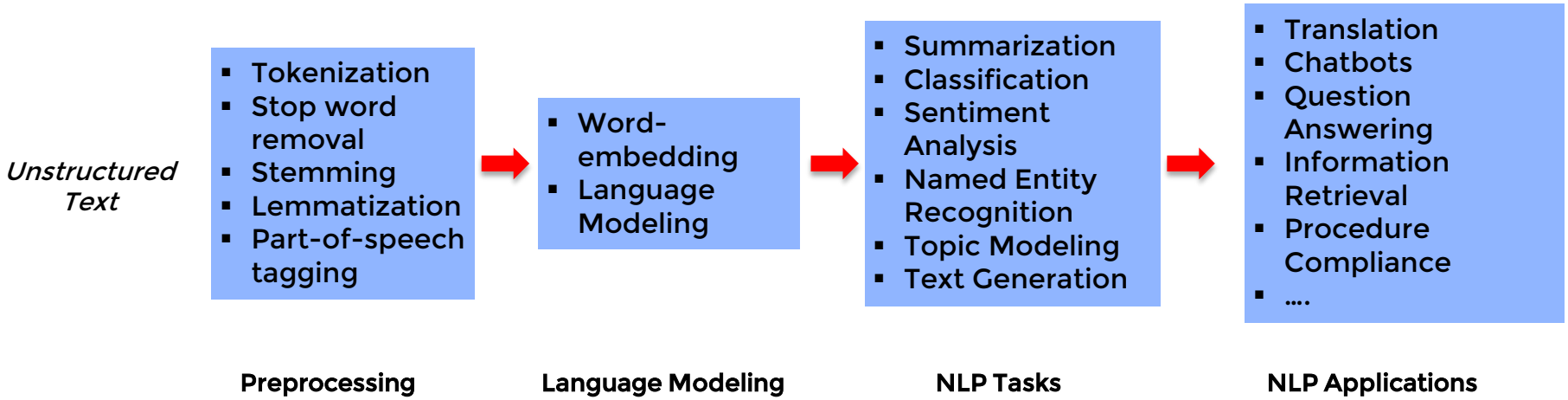


ChatGPT

Expert's Opinion



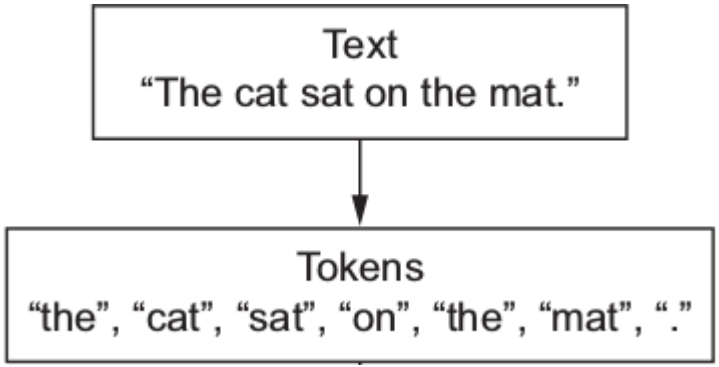
Workflow


Unstructured Text
Tokenization
Stop word removal
Stemming
Lemmatization
Part-of-speech tagging
Word-embedding
Language Modeling
Summarization
Classification
Sentiment Analysis
Named Entity Recognition
Topic Modeling
Text Generation
Translation
Chatbots
Question Answering
Information Retrieval
Procedure Compliance
....
Preprocessing
Language Modeling
NLP Tasks
NLP Applications

Preprocessing

Tokenization

- To break text down into a smaller units (e.g. words, sentences). Each of the smaller unit is called tokens



- The tokens could be words, numbers or punctuation marks. In tokenization, smaller units are created by locating word boundaries.

Preprocessing

Stop word removal

- The words which are generally filtered out before processing a text are called “**stop words**”.

- These are actually the most common words in any language (like articles, prepositions, pronouns, conjunctions, etc) and does not add much information to the text.
  - Examples: **“the”, “a”, “an”, “so”, “what”**

- *Definition of stop words is application dependent*

Preprocessing

Stemming & Lemmatization

- The goal of both stemming and lemmatization is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form.

  car, cars, car's, cars' ➔ car

  am, are, is ➔be;

- **Stemming** is the process of reducing words to their word stem. A "**stem**" is the part of a word that remains after the removal of all affixes.
  - the stem for the words "touched", "touch." and "touching" is "touch".

- **Lemmatization** is the process to return to the base or dictionary form of a word, which is known as "**lemma"**

Preprocessing

Part of the speech (POS) tagging & Parsing

- POS is to categorize words/phrases in a text (corpus) in correspondence with a particular part of speech, depending on the definition of the word and its context.
- Parsing - Formal analysis of a sentence into its constituents, which results in a parse tree showing their syntactic relation to one another in visual form

Why	not	tell	someone	?
adverb	adverb	verb	noun	punctuation mark, sentence closer

The thief robbed the apartment

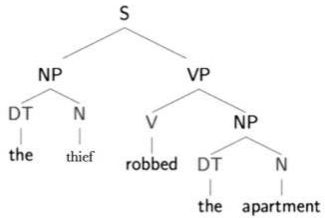
1) Part of speech

N = noun
V = verb
DT = determiner

2) Phrases

Noun Phrases: : "the thief", "the apartment"
Verb Phrases: "robbed the apartment"
Sentence: "the burglar robbed the apartment"

3) Relationships



POS applications:

- Assumption about semantics
- Name entity recognition
- Co-reference resolution
- Speech recognition

"I voted for Trump because he was most aligned with my values", John said.
The original sentence

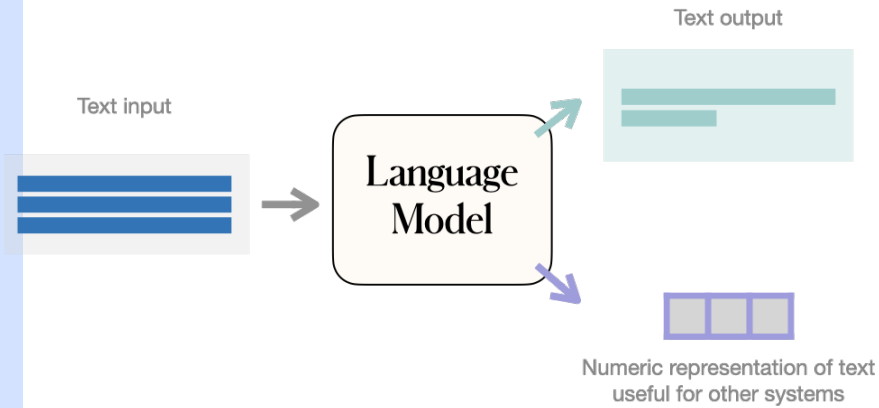
"John voted for Trump because Trump was most aligned with John's values", John said.
The sentence with resolved coreferences

Language Modeling

What is it?

- Language Modeling (LM) is the development of probabilistic models ***that are able to predict the next word in the sequence given the words that precede it***.

- LM learns the probability of **word occurrence** based on examples of text. *Simpler models may look at a context of a short sequence of words, whereas larger models may work at the level of sentences or paragraphs.*

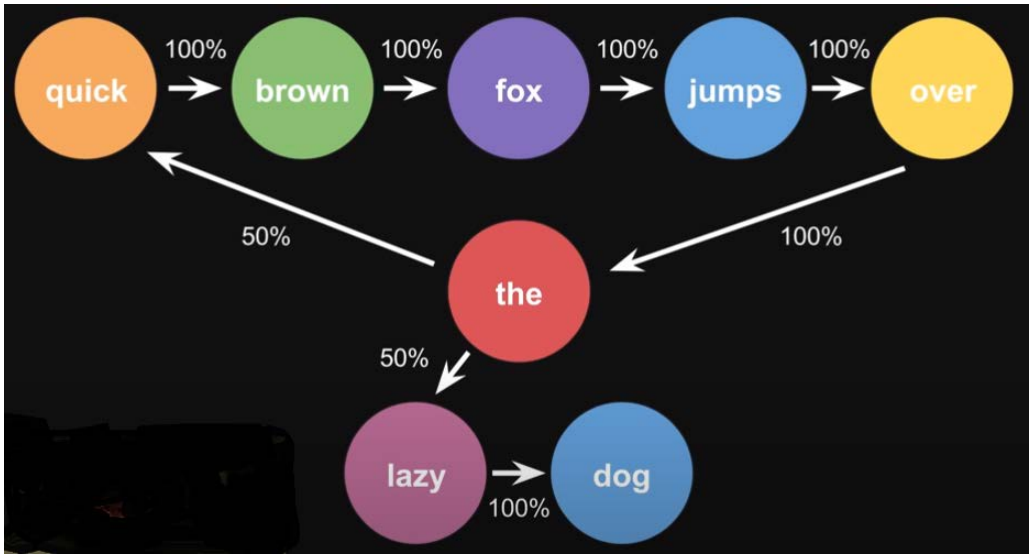


Language Modeling

Statistical language models (SLM)

Markov Chains

$$\Pr(X_{n+1} = x \mid X_n = x_n)$$



Example: “the quick brown fox jumps over the lazy dog”

Language Modeling

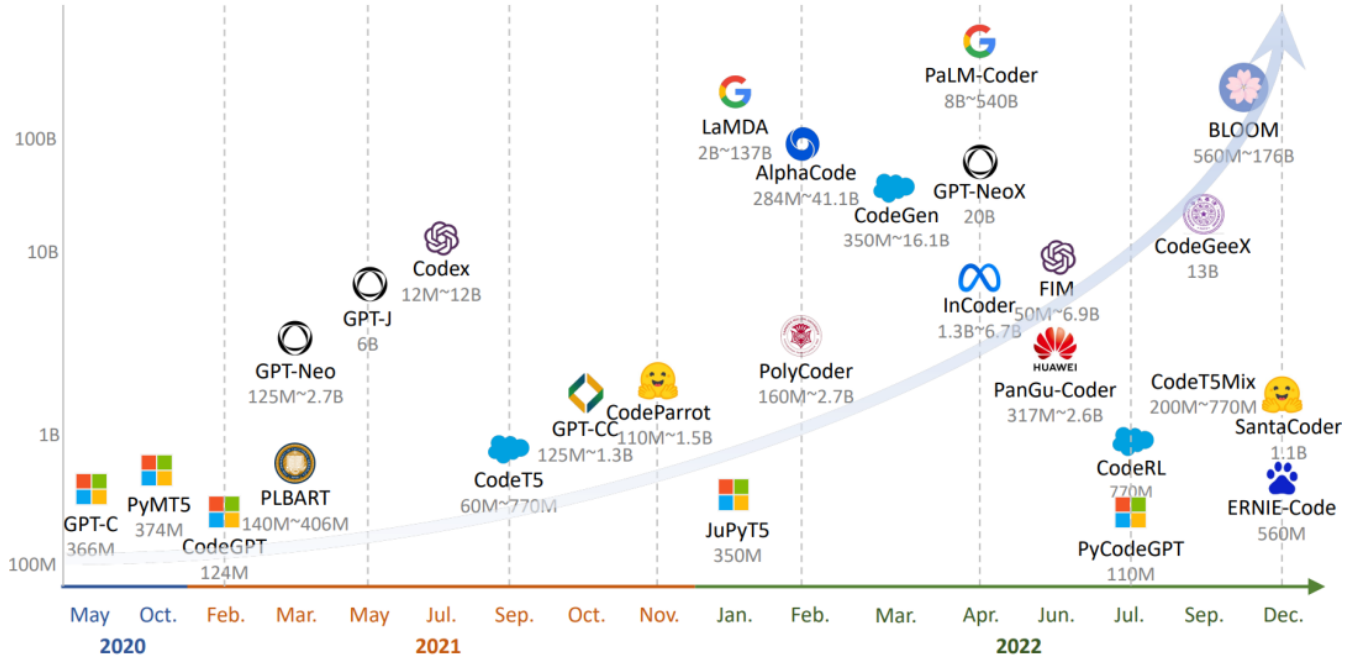
Statistical language models (SLM)

This is Big Data AI Book

Language Modeling

Neural Language Models (NLM)

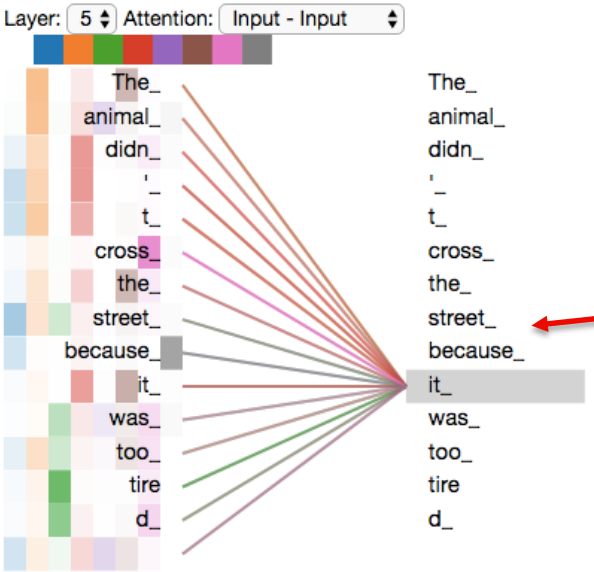
- Make use of **Neural networks**.
  - Most popular one is "**Transformer**"



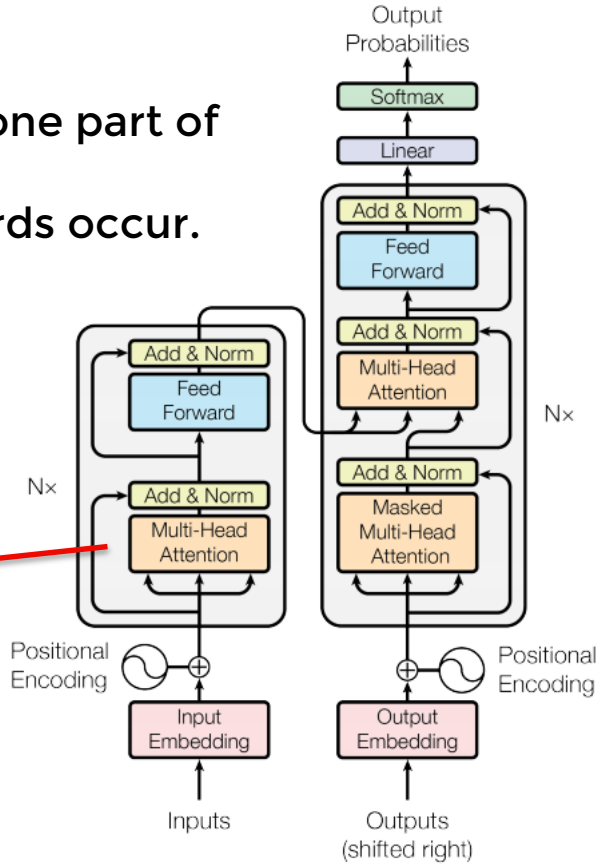
Language Modeling

Neural Language Models - Transformer

- Self-attention: assigns a weight (importance) to one part of an input against the rest of the input
- Positional encoding: the order in which input words occur.



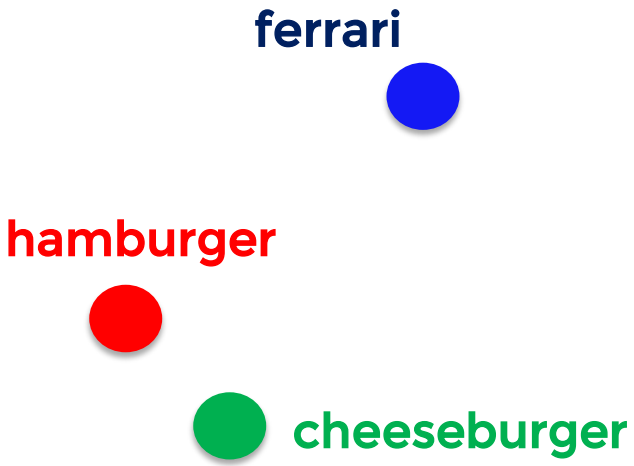
Attention

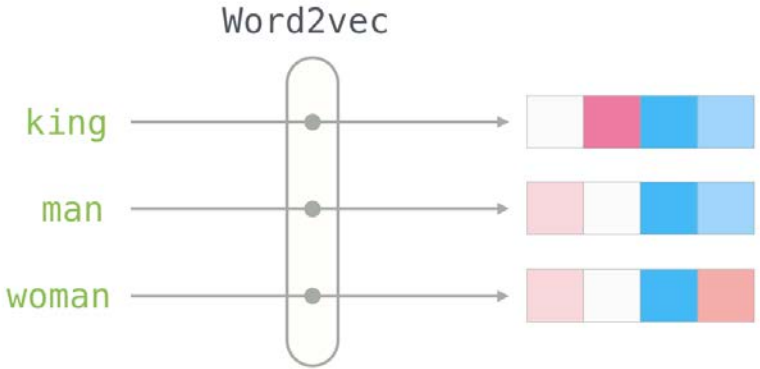


Word Embeddings

Language Modeling(word/context) = Word Embeddings

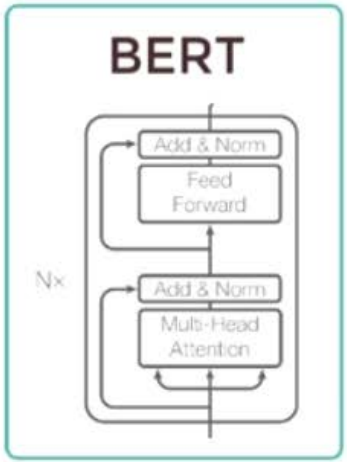
- A **word embedding** is a representation of a word, typically consisting of a real-valued vector that **encodes the meaning of the word** in such a way that words that are closer in the vector space are expected to be similar in meaning.





Transformer - 1

BERT – Bidirectional Encoder Representation from Transformers



Problems to Solve

- Neural Machine Translation
- Question Answering
- Sentiment Analysis
- Text summarization

Needs Language understanding

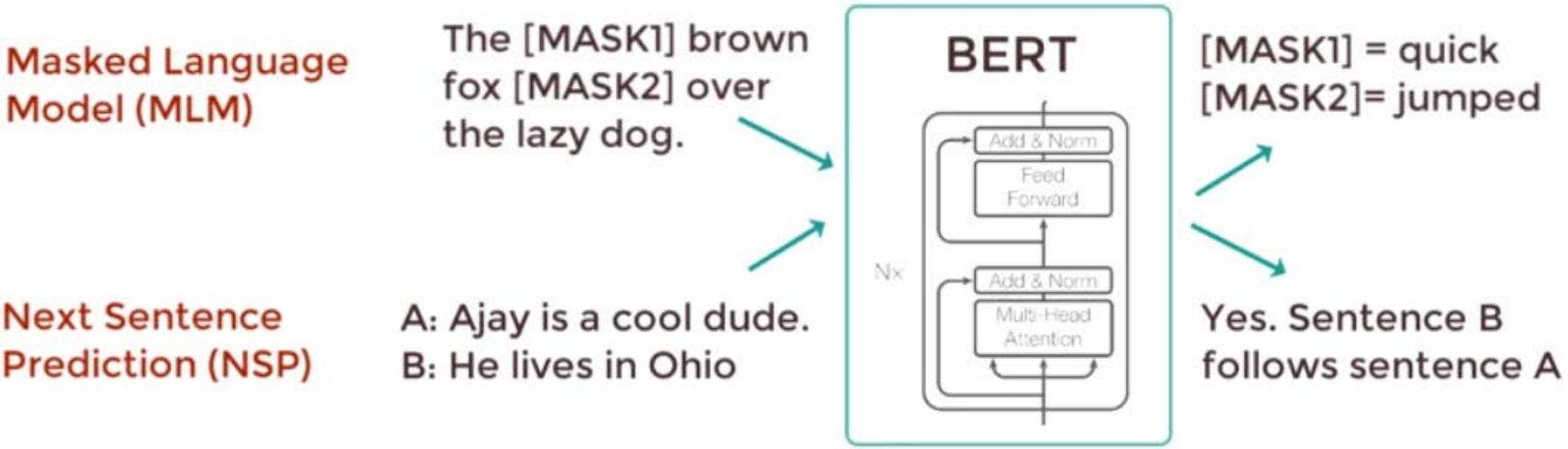
How to solve Problems (BERT Training)

- Pretrain BERT to understand langauge
- Fine tune BERT to learn specific task

Transformer - 1

BERT – Bidirectional Encoder Representation from Transformers

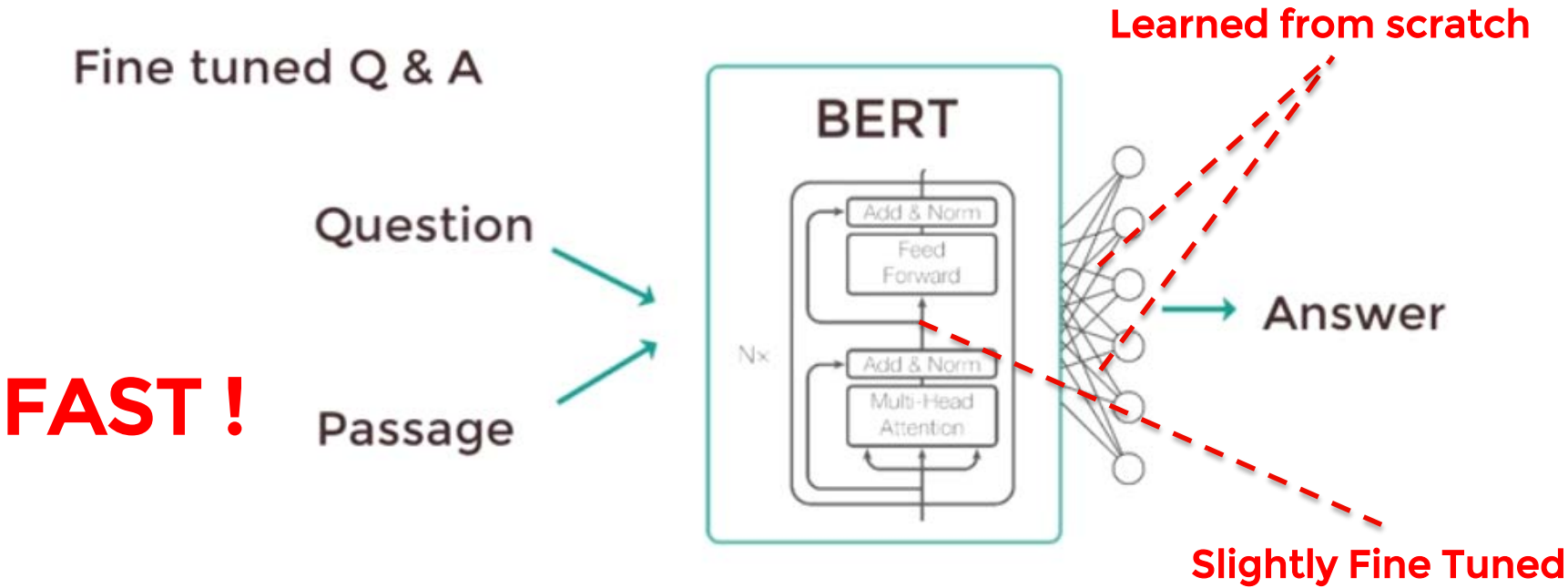
Pretraining : “What is the language? What is the context?”



Transformer - 1

BERT – Bidirectional Encoder Representation from Transformers

Fine Tuning : "What to use language for specific task?"



Transformer - 1

BERT – Bidirectional Encoder Representation from Transformers

	BERT	RoBERTa	DistilBERT	XLNet
Size (millions)	**Base**: 110 **Large**: 340	**Base**: 110 **Large**: 340	**Base:** 66	**Base**: ~110 **Large**: ~340
Training Time	**Base**: 8 x V100 x 12 days* **Large:** 64 TPU Chips x 4 days (or 280 x V100 x 1 days*)	**Large:** 1024 x V100 x 1 day; 4-5 times more than BERT.	**Base**: 8 x V100 x 3.5 days; 4 times less than BERT.	**Large:** 512 TPU Chips x 2.5 days; 5 times more than BERT.
Performance	Outperforms state-of-the-art in Oct 2018	2-20% improvement over BERT	3% degradation from BERT	2-15% improvement over BERT
Data	16 GB BERT data (Books Corpus + Wikipedia). 3.3 Billion words.	160 GB (16 GB BERT data + 144 GB additional)	16 GB BERT data. 3.3 Billion words.	**Base**: 16 GB BERT data **Large**: 113 GB (16 GB BERT data + 97 GB additional). 33 Billion words.
Method	BERT (Bidirectional Transformer with MLM and NSP)	BERT without NSP**	BERT Distillation	Bidirectional Transformer with Permutation based modeling

Transformer - 2

GPT - Generative Pretrained Transformer

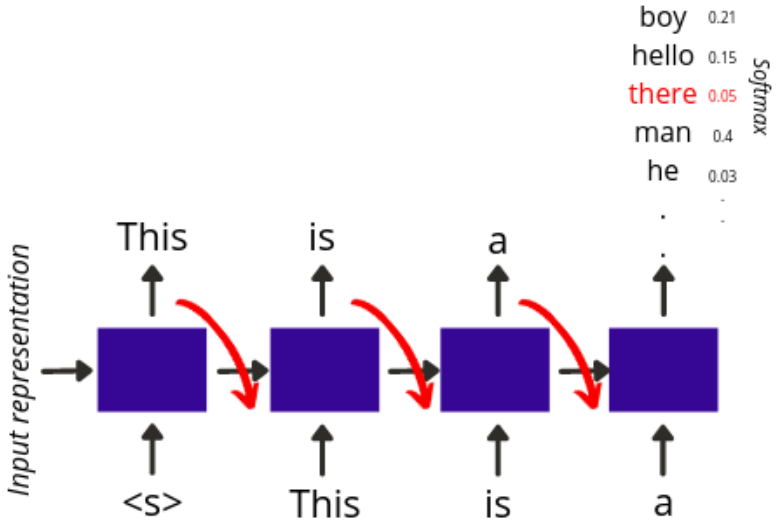
A process of generating text with the goal of appearing indistinguishable to human-written text

ChatGPT broke the Turing test – the race is on for new ways to assess AI (Nature, 25 July 2023)

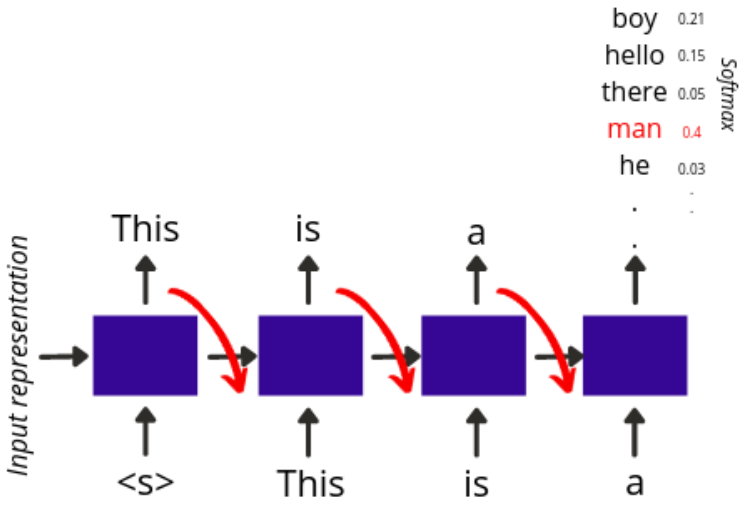
V1: 117 Million Parameters
V2: 1.5 Billion Parameters
V3: 175 Billion Parameters
V3.5: ChatGPT

Text generation

common strategies: statistical correlations between words



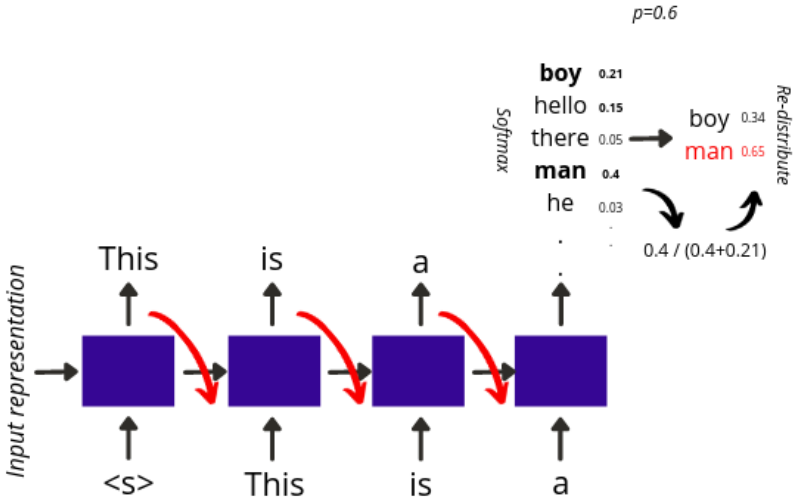
Randomly Sampling



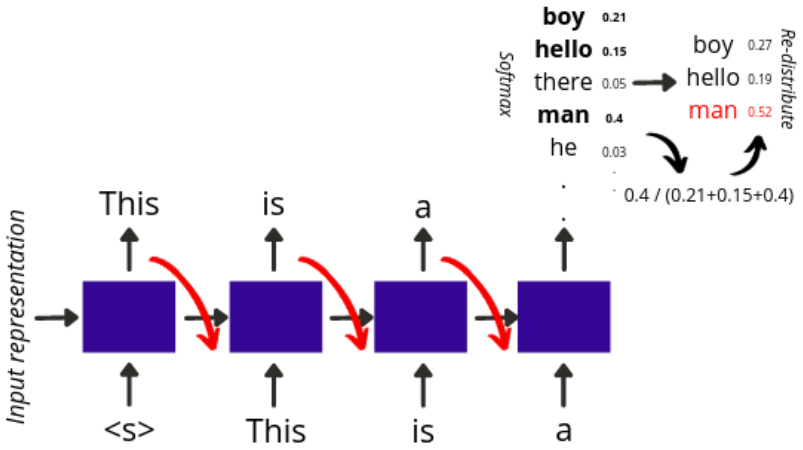
Greedy Decoding

Text generation

common strategies: statistical correlations between words



Top-p or Nucleus Sampling (threshold)



Top-k Sampling (random)

Prompt Engineering

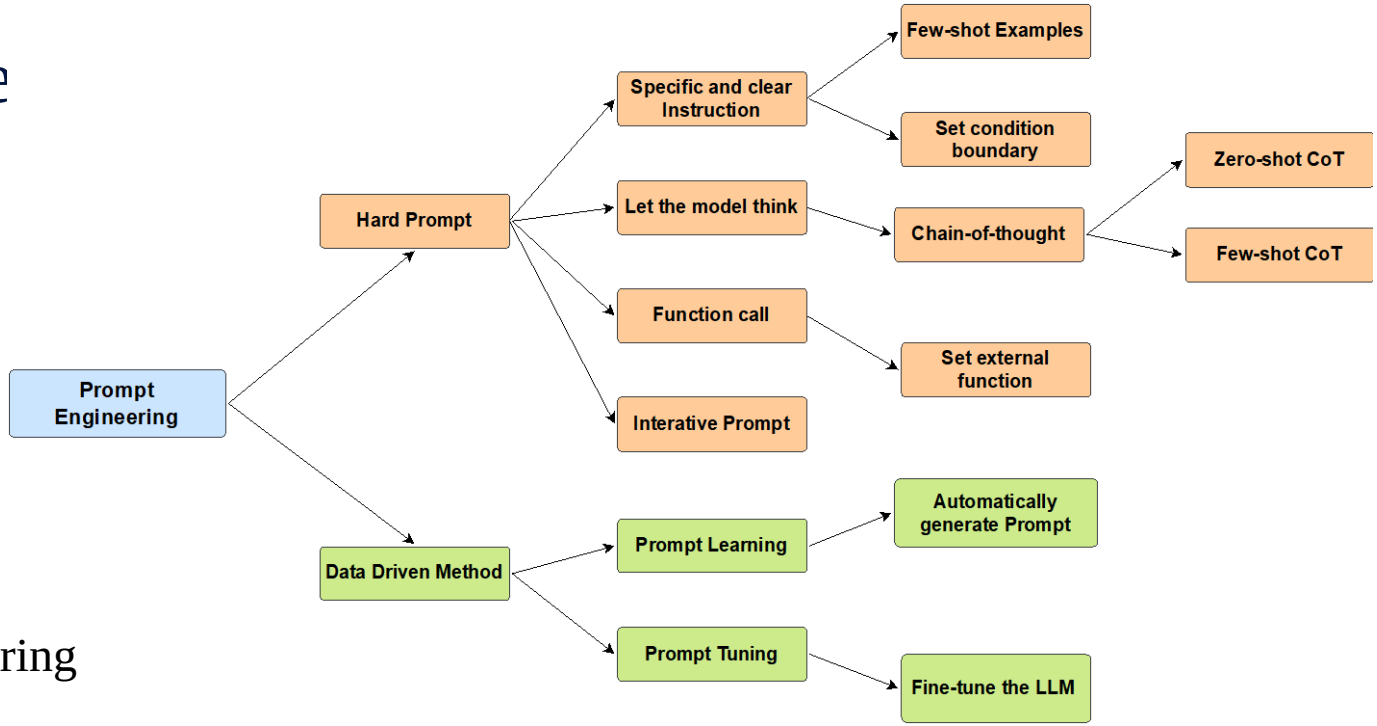
UNIVERSITY
OF WOLLONGONG
AUSTRALIA

Prompt Engineering

Design and optimize prompts for LM, such as the GPT or Llama, to generate the specific output required by the user.

- Prompts can be questions, commands, statements even token or parameters
- Possible to solve problems without directly re-training data.

Prompt Engine



Hard Prompt

- a direct input string

Data Driven Method

- automatically optimize the prompt through the data learning process or fine-tune the model

Hard Prompt

Hand-made, predefined text (templates)

Techniques:

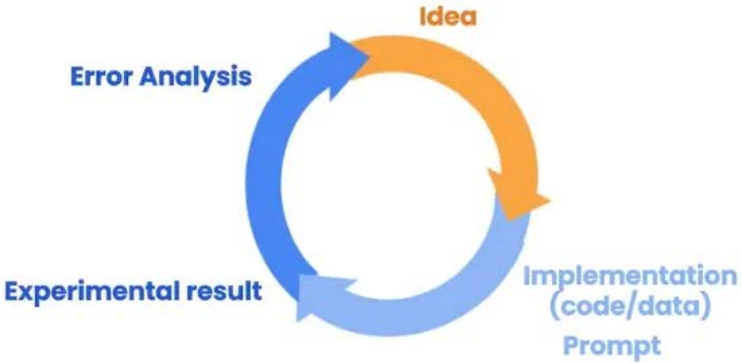
- Specific and clear instructions
  - Structured input or output
  - Set condition boundary
  - Few-shot example
- Let the model think
  - Chain-of-thought (Zero-shot or Few-shot CoT)
- Function call
  - Call external function or API
- Iterative Prompt
  - Iterative trial and error and experimentation to get a prompt that is appropriate for each domain and task

- To categorize movie reviews, the prompt could be: "Was this movie reviewed positively or negatively?" + [Text of film review].

Hard Prompt

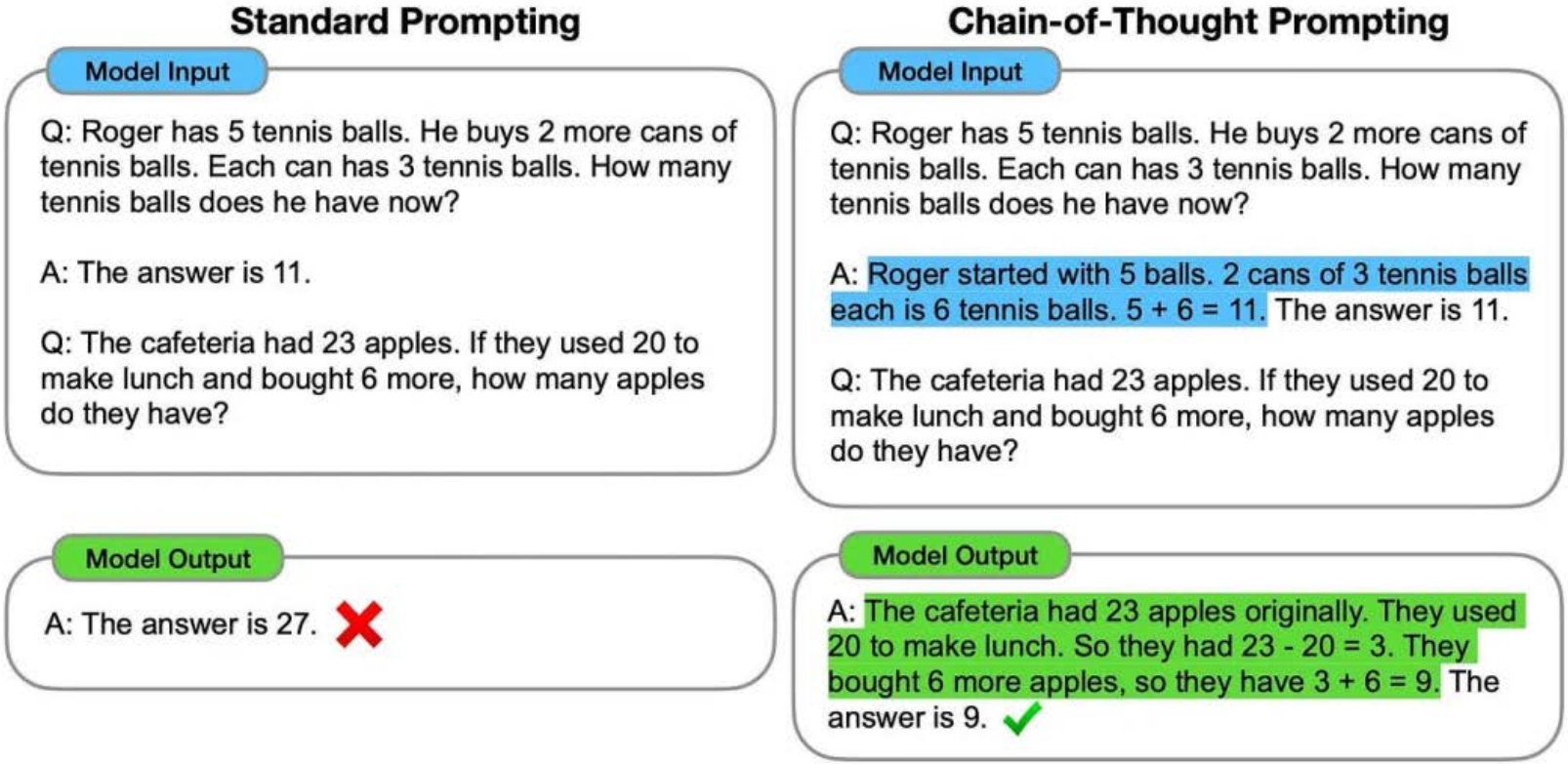
Six elements

1. Task (generate, drive, write, analyse)
2. Context (background)
3. Examples
4. Roles (similar to human experience)
5. Format
6. Tone



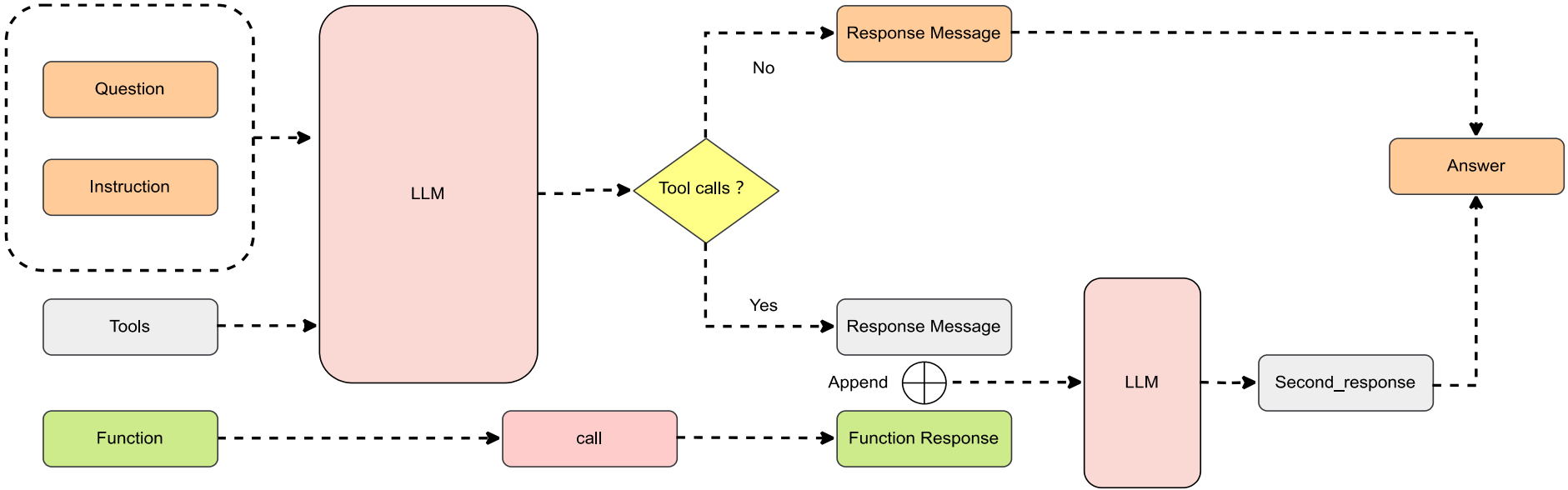
Hard Prompt

Example of CoT



Hard Prompt

Function call



Hard Prompt

Iterative Prompt

Big to small chunks

Prompt 0

Question: It takes Amy 4 minutes to climb to the top of a slide. It takes her 1 minute to slide down. The water slide closes in 15 minutes. How many times can she slide before it closes?
<LM>
Answer: To solve "How many times can she slide before it closes?", we need to first solve: "How long does each trip take?"
</LM>

Prompt 1

It takes Amy 4 minutes to climb to the top of a slide. It takes her 1 minute to slide down. The water slide closes in 15 minutes.
Subquestion 1: How long does each trip take?
<LM>
Answer 1: It takes Amy 4 minutes to climb and 1 minute to slide down. $4 + 1 = 5$. So each trip takes 5 minutes.
</LM>

Prompt 2

It takes Amy 4 minutes to climb to the top of a slide. It takes her 1 minute to slide down. The slide closes in 15 minutes.
Subquestion 1: How long does each trip take?
Answer 1: It takes Amy 4 minutes to climb and 1 minute to slide down. $4 + 1 = 5$. So each trip takes 5 minutes.
Subquestion 2: How many times can she slide before it closes?
<LM>
Answer 2: The water slide closes in 15 minutes. Each trip takes 5 minutes. So Amy can slide $15 \div 5 = 3$ times before it closes.
</LM>

Data driven method

Prompt learning

- Emphasis on automatic discovery or optimization of the most effective prompts through the learning process.
  - For example, automatic generation of efficient prompts for domain-specific questions improves the accuracy of question answering systems.

Prompt tuning

- Adjust a few parameters of the model to better respond to a given prompt
  - For example, fine-tuning the model to fit specific types of sentiment analysis tasks and improving the model's ability to recognize emotions through precise adjustments.

Prompt learning

Process:

(1) Ask the model to generate a candidate set of prompts

(2) Filter the candidate set based on a selected scoring function

$$\rho^* = \arg\max_\rho \mathbb{E}_{(x,y)\in\mathcal{D}_{\text{train}}}[f(\rho, x, y)]$$

APE: Larger Language models are human-level prompt engineers

Conclusion

Promising outcomes and Challenges

Promising outcomes

- Neural language models enabling many applications
  - ✓ BERT and its variants (Google)
  - ✓ GPT-x and its fine-tuned model, e.g. ChatGPT (OpenAI)
- Turning language computable
  - ✓ Inference via contemporary ML and DL
  - ✓ Fusion of different types of information
- 77% of Businesses Using NLP Expect to Increase Investment (recent survey)

Some Challenges

- Ambiguity
- Context dependence
- Out-of-Vocabulary words
- Annotation bias
- Multi-lingual NLP
- Large number of parameters of LMs
- 67% of businesses with NLP models in production for 5+ years still deal with accuracy challenges (recent survey)

Thank you!

UNIVERSITY
OF WOLLONGONG
AUSTRALIA