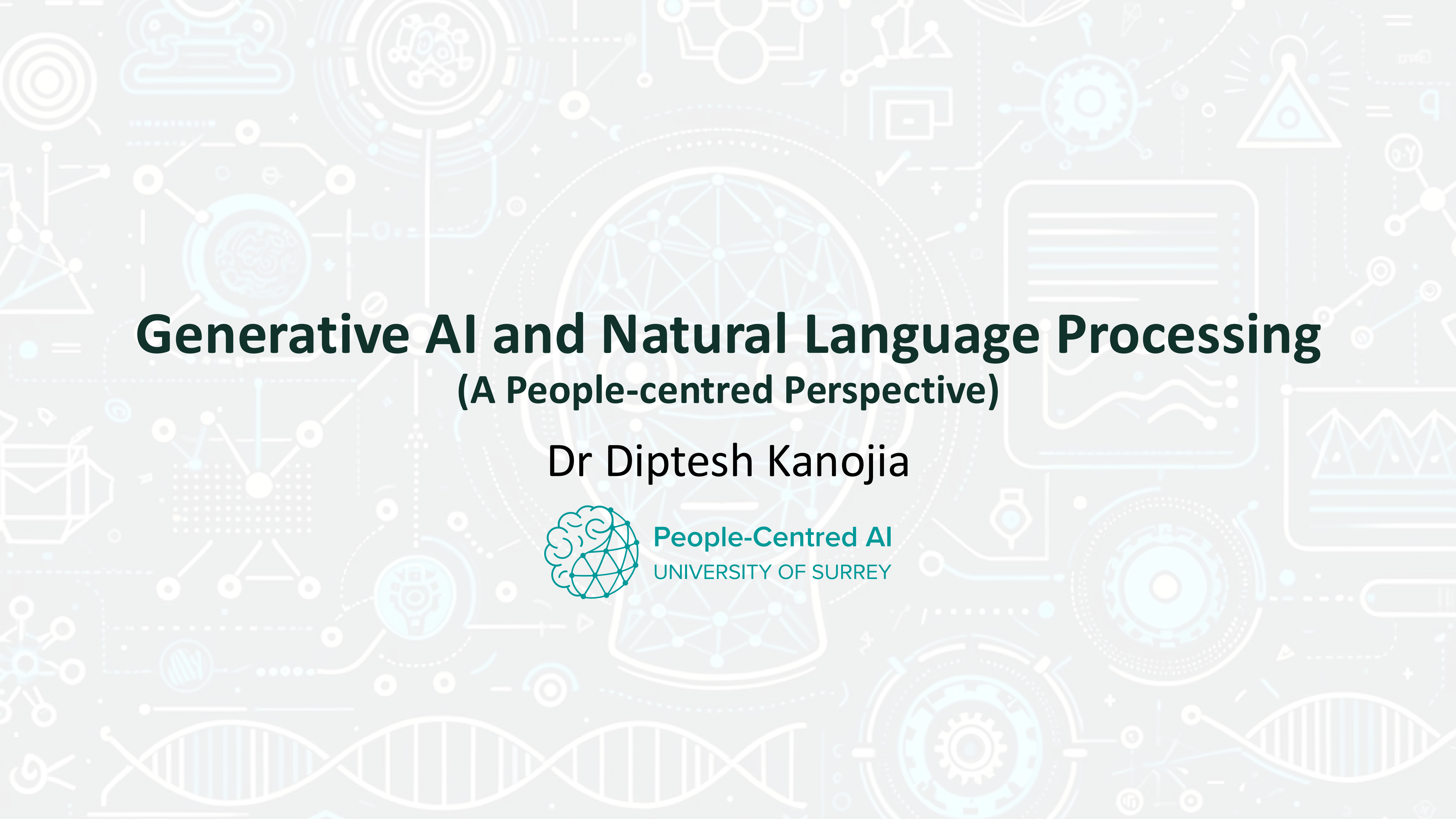


Generative AI and Natural Language Processing

(A People-centred Perspective)

Dr Diptesh Kanojia

People-Centred AI
UNIVERSITY OF SURREY

Agenda

Introduction to Generative AI and NLP

Understanding Large Language Models

The Evolution of Language Models Leading to ChatGPT

Recent Advancements in Language Processing

Challenges and Risks

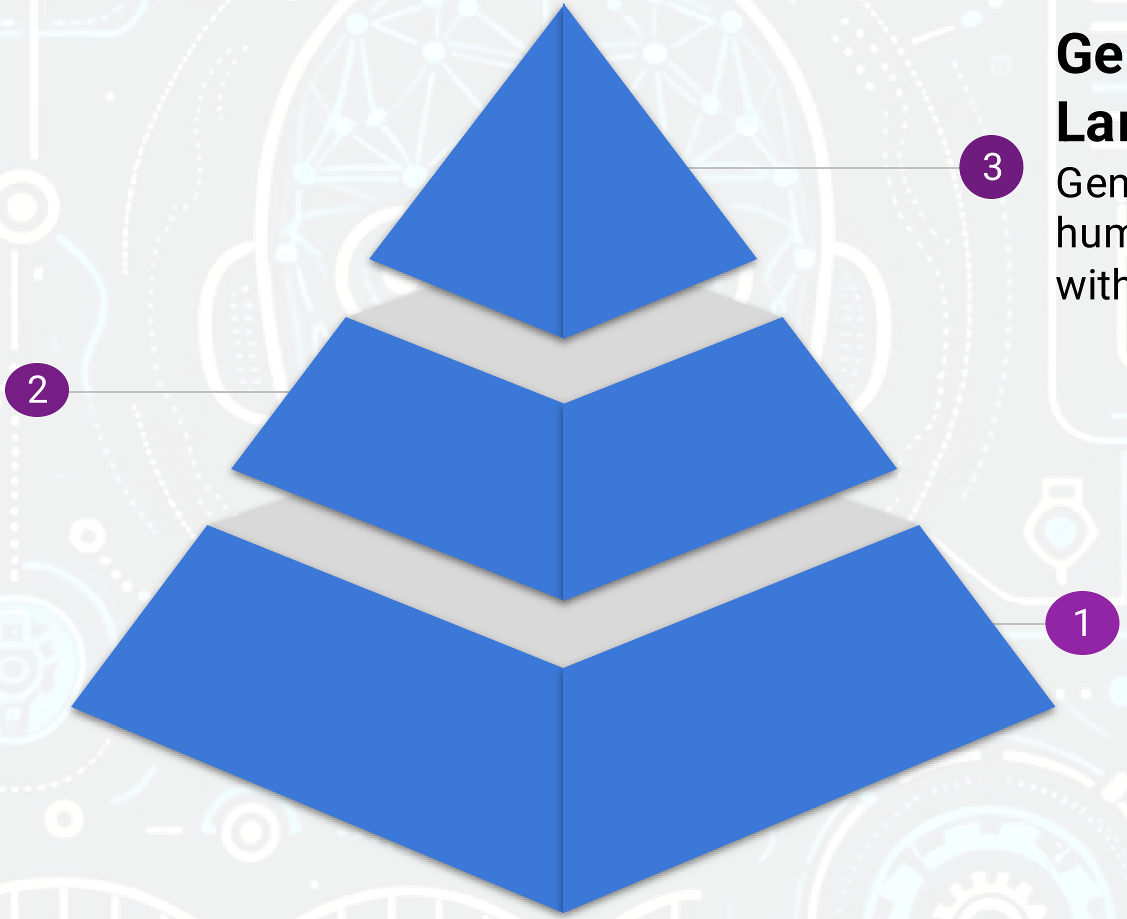
A People-Centred Approach to NLP

Spotlight on Initiatives

Generative Artificial Intelligence (GenAI)

- Language and GenAI
  - NLP-driven advancements in the field of Artificial Intelligence
  - “Transformers” leading innovation in Computer Vision and Speech, alike.
- Commonalities
  - Input (text, image, video, speech) is always converted to mathematical representation of the input
- **Language Modelling -** build a general world view for concepts within a language, inside a computational model, which can perform certain tasks.

Natural Language Processing (NLP)



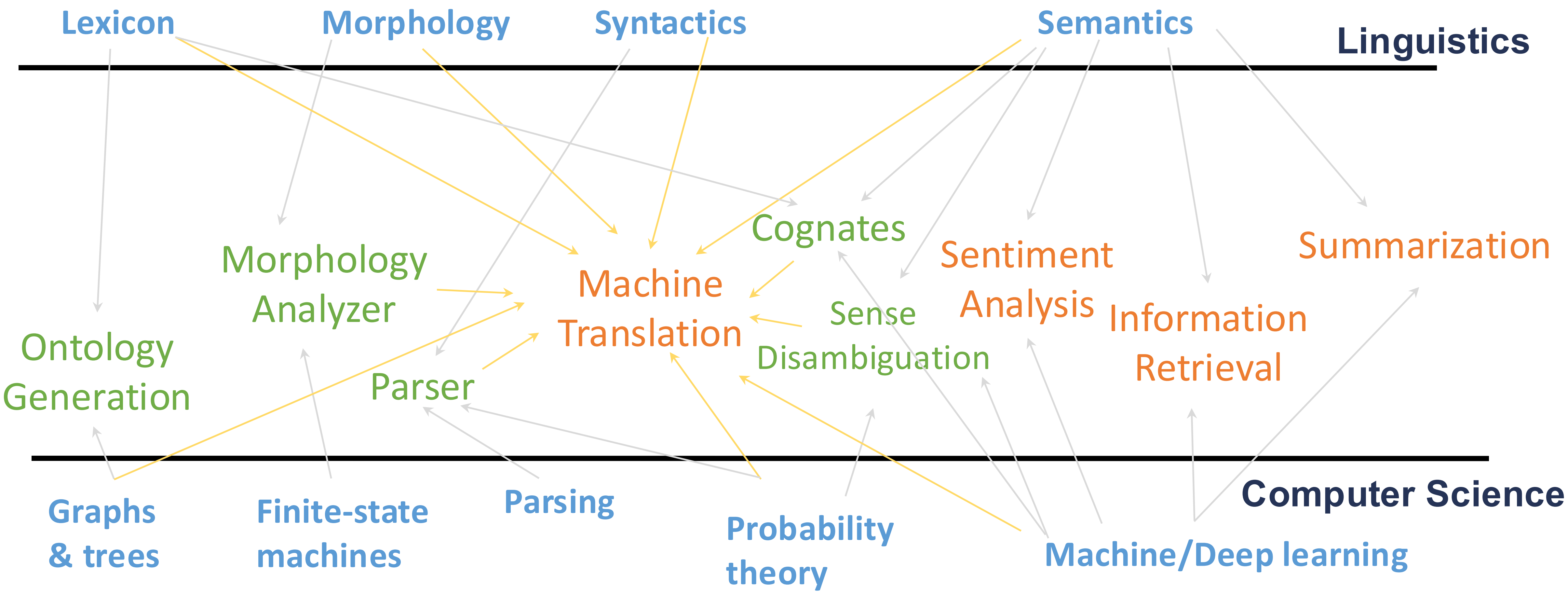
Generate Human Language
Generation of understandable human language to interface with people.

Analyse Human Language
Textual analytics, extraction, and retrieval to analyze the information present in human language.

"Understand" Human Language

A key goal of NLP research is to ensure that machines understand human language.

Natural Language Processing *vs. Computational Linguistics*



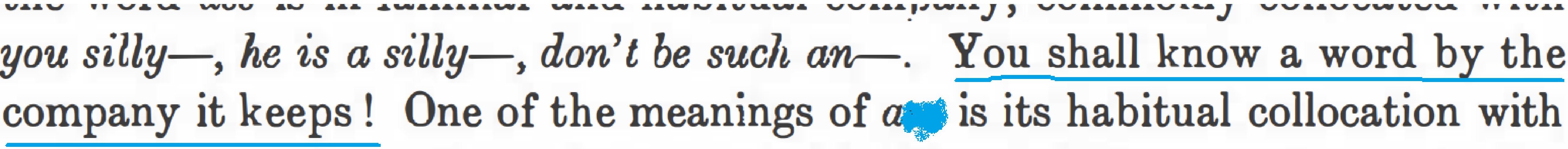
Lexicon
Morphology
Syntactics
Semantics
Linguistics

Cognates
Morphology Analyzer
Machine Translation
Sentiment Analysis
Summarization
Ontology Generation
Sense Disambiguation
Information Retrieval
Parser

Computer Science
Graphs & trees
Finite-state machines
Parsing
Probability theory
Machine/Deep learning

Base slide credits: Prof Pushpak Bhattacharyya

How it started?



you silly—, he is a silly—, don't be such an—. You shall know a word by the company it keeps! One of the meanings of a[redacted] is its habitual collocation with

R, Firth J. (1957). "A synopsis of linguistic theory, 1930-1955". *Studies in Linguistic Analysis*.

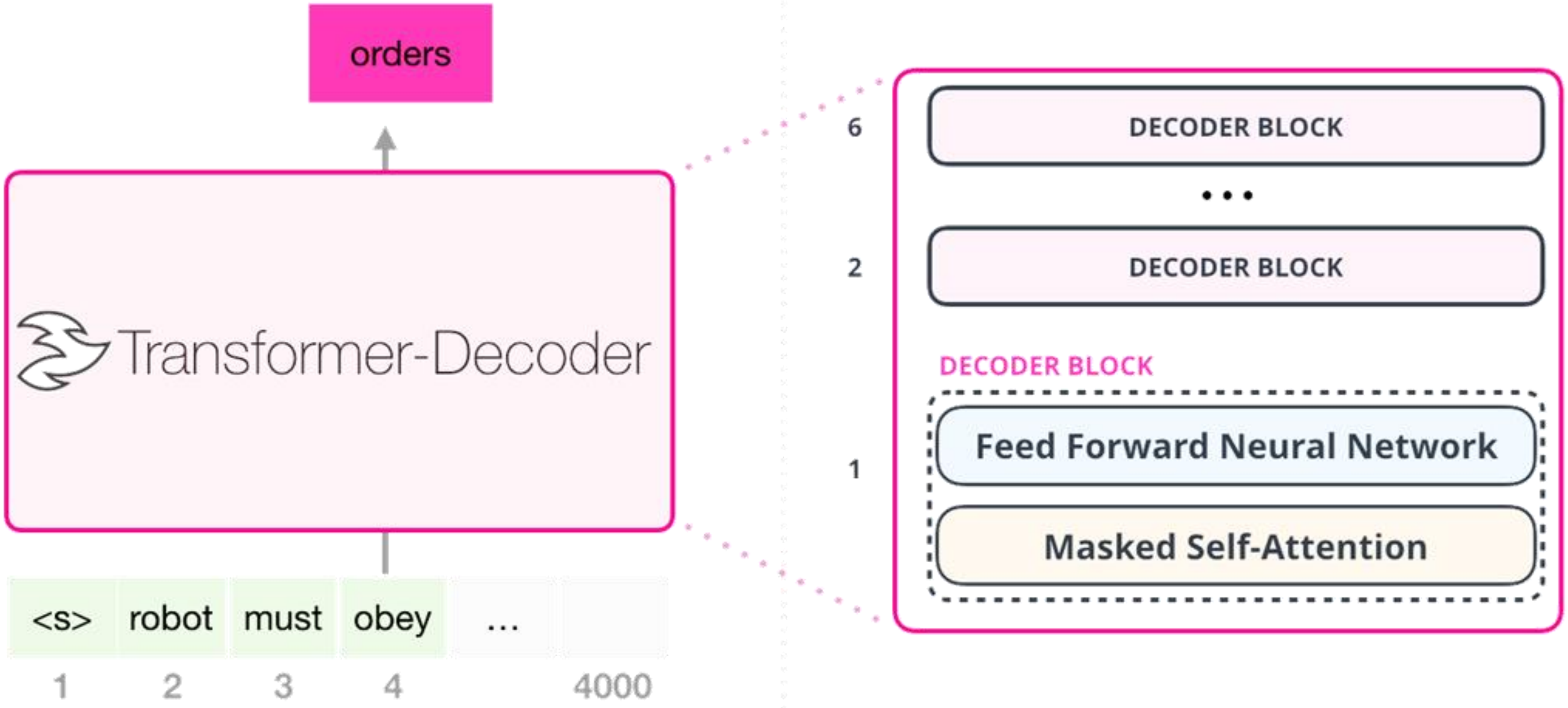
The 'context' of a word is important to decipher its meaning or its sense.

Inspired generations of research on Word Sense Disambiguation, Language Modeling, and Machine Translation.

Many approaches from statistical modeling of language are grounded on this principle.

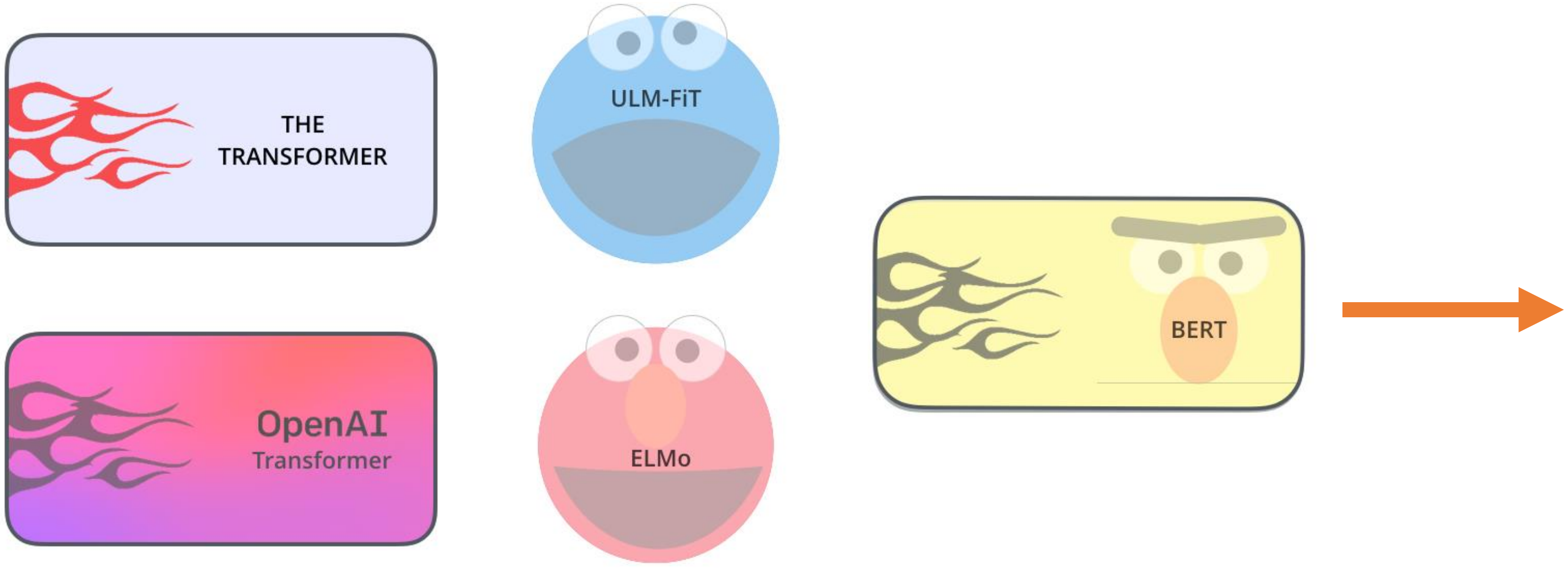
Language Modeling: The core of ChatGPT

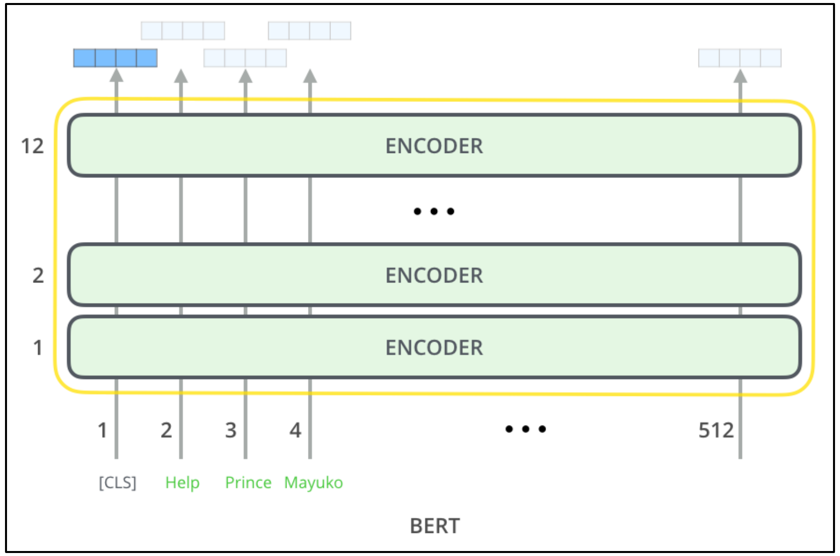
Task: iterative next word prediction



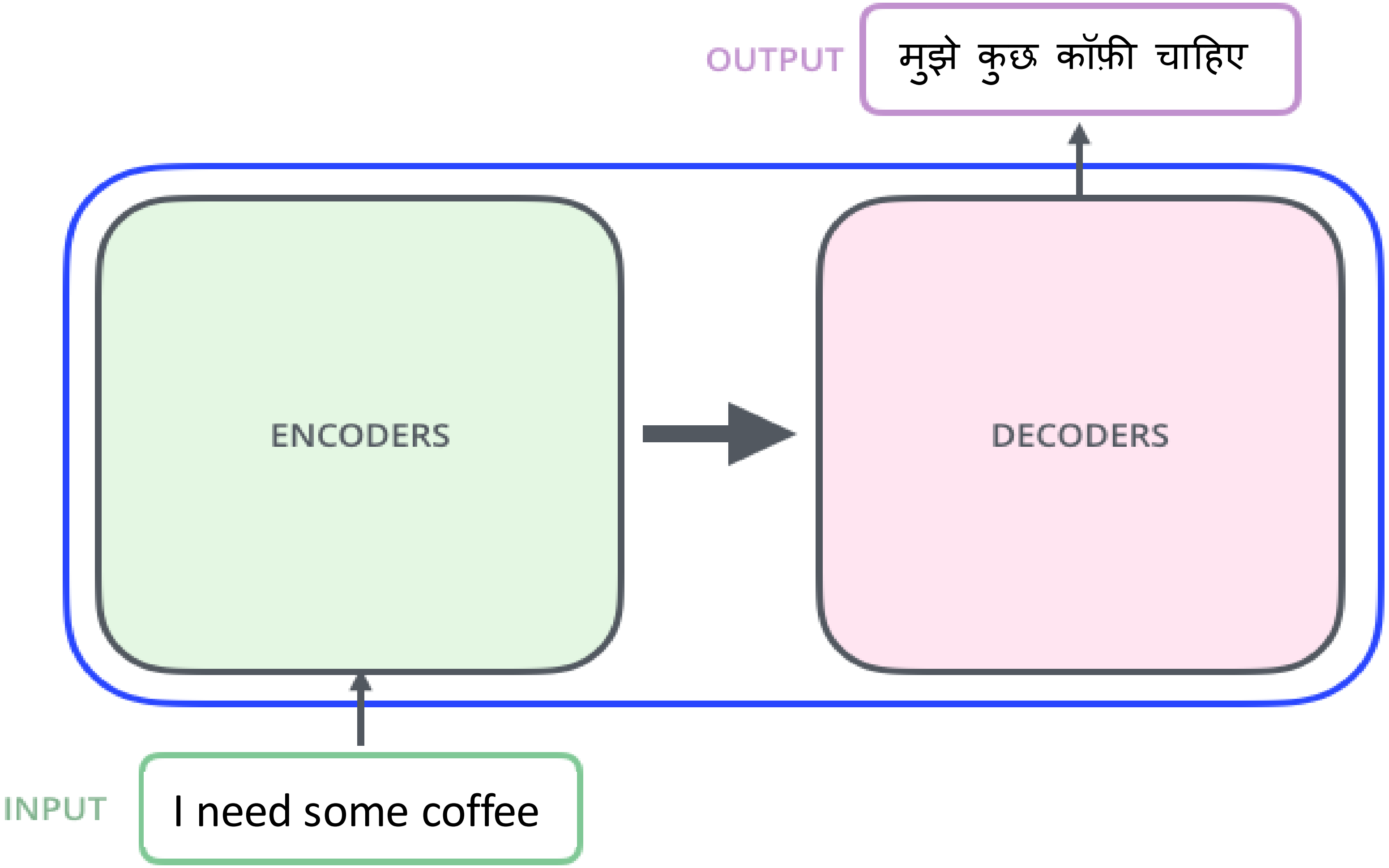
From *ElMo* to *BERT*: Leap in Language Modeling

- Attention is all you need (Vaswani et. al., 2017) proposed the use of self-attention in language encoding and decoding processes.

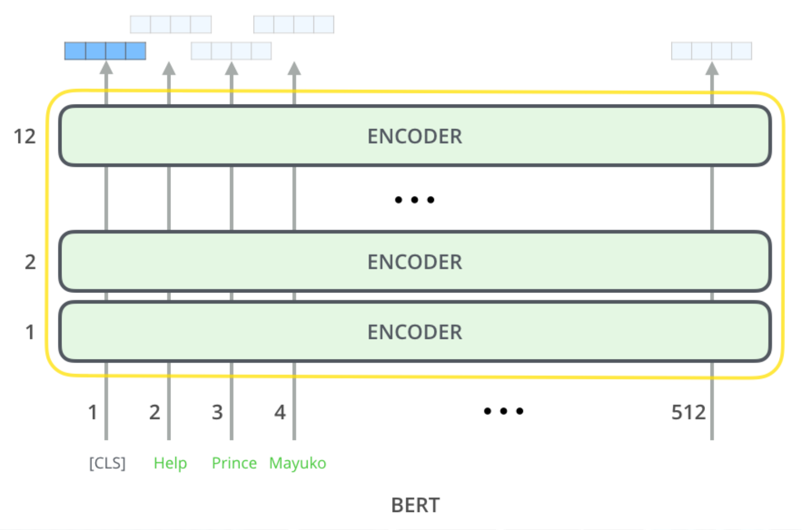


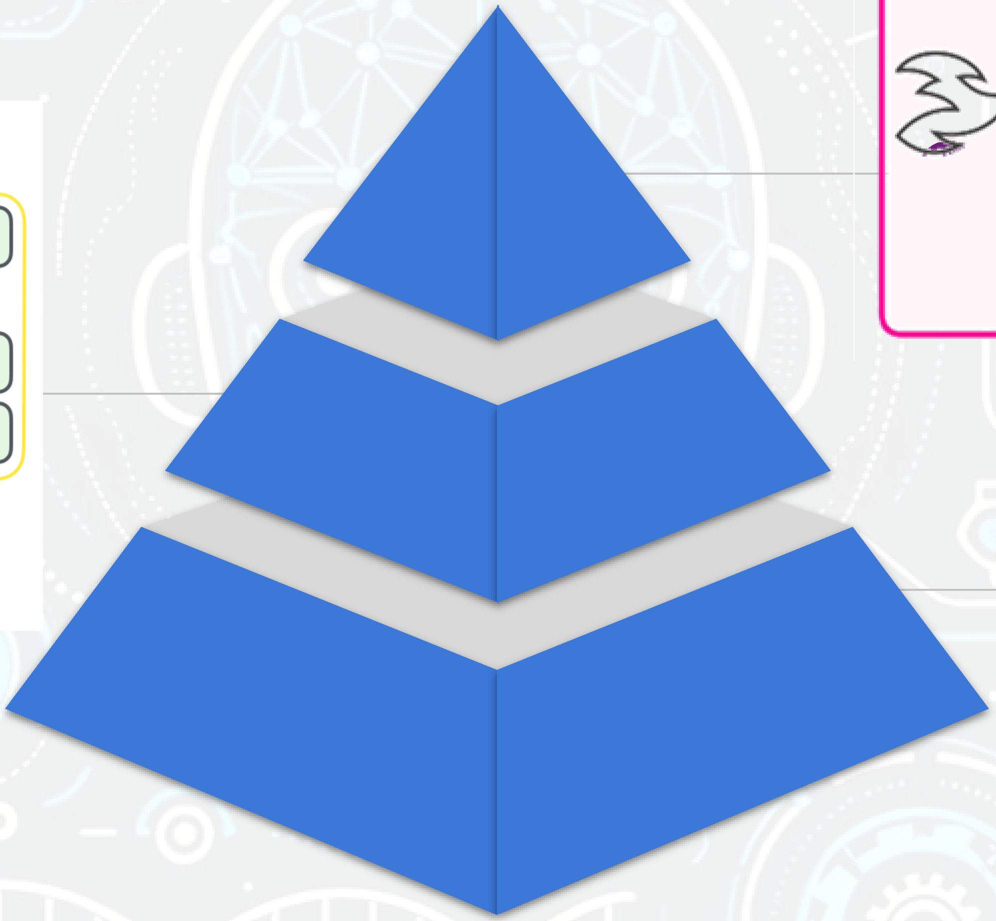


Encoder *vs.* Decoder



Natural Language Processing (NLP)





Transformer-Decoder

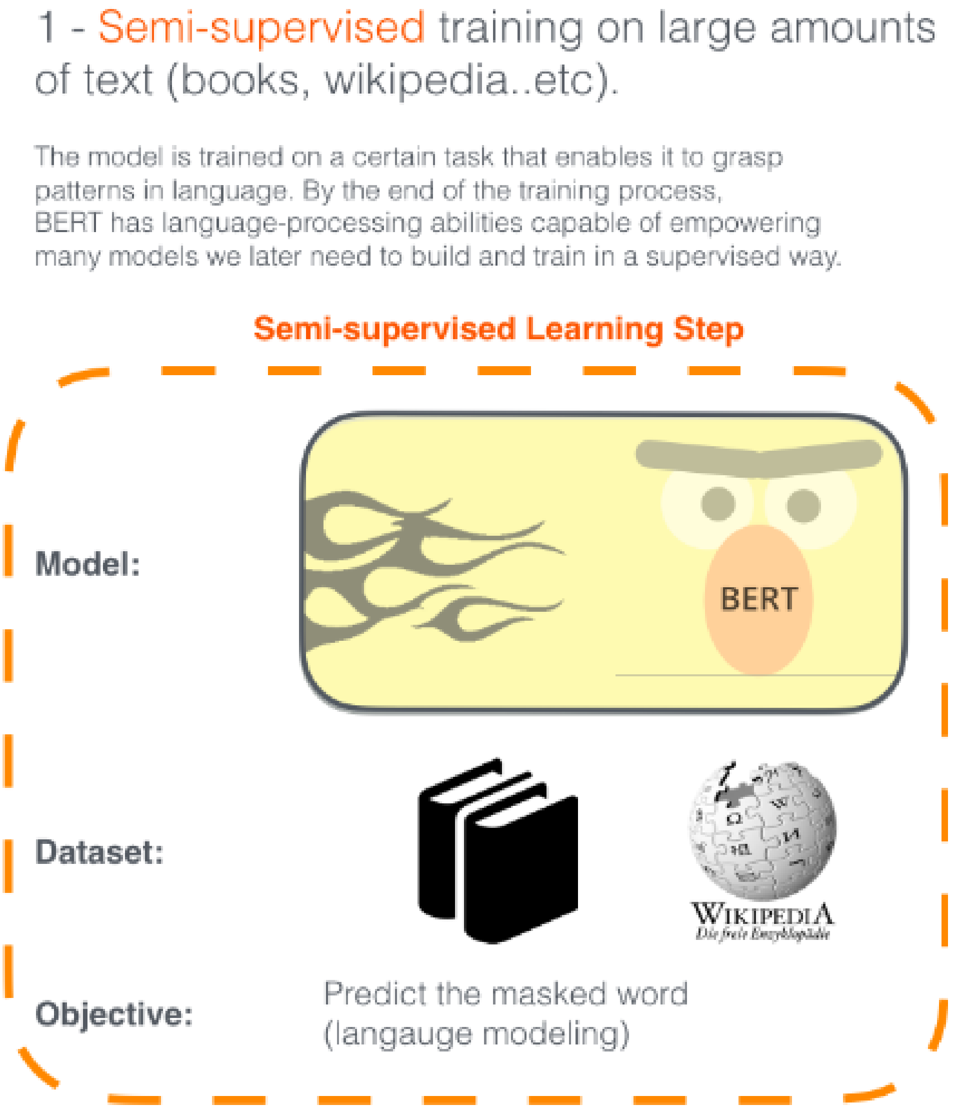
'Understand' Human Language

1

A key goal of NLP research is to ensure that machines understand human language.

Engineering the backbone of NLP: Pre-train

- A novel ‘pre-training’ objective for language modeling
    - Fill in the [MASK] given a sentence like:
        - I love making slides at 11 PM in the [MASK]
        - Why would I schedule a [MASK] during a busy semester?

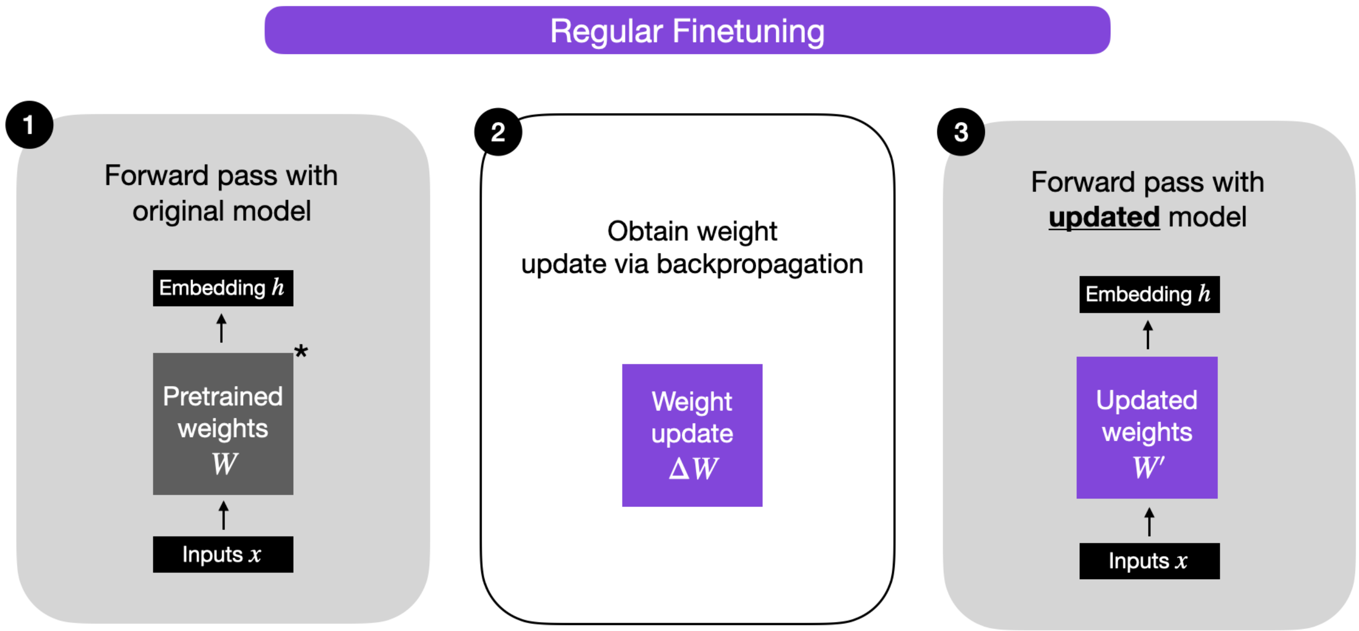


1 - Semi-supervised training on large amounts of text (books, wikipedia..etc).

The model is trained on a certain task that enables it to grasp patterns in language. By the end of the training process, BERT has language-processing abilities capable of empowering many models we later need to build and train in a supervised way.

Semi-supervised Learning Step
Model:
BERT
Dataset:
WIKIPEDIA
Die freie Enzyklopädie
Objective:
Predict the masked word (langauge modeling)

Engineering the backbone of NLP: Transfer Learn



Regular Finetuning

1 Forward pass with original model

Embedding h ↑ Pretrained weights W* ↑ Inputs x

2 Obtain weight update via backpropagation

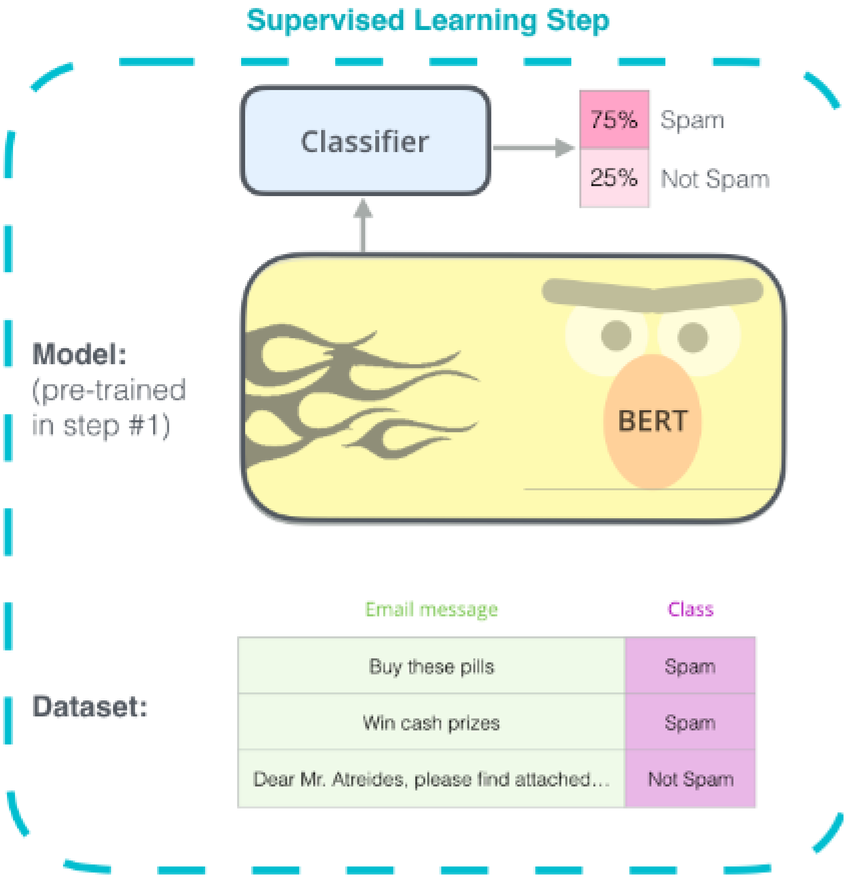
Weight update ΔW

3 Forward pass with **updated** model

Embedding h ↑ Updated weights W' ↑ Inputs x

* The pretrained model could be any LLM, e.g., an encoder-style LLM (like BERT) or a generative decoder-style LLM (like GPT)

2 - Supervised training on a specific task with a labeled dataset.



Supervised Learning Step

Model: (pre-trained in step #1)

Dataset:

Email message	Class
Buy these pills	Spam
Win cash prizes	Spam
Dear Mr. Atreides, please find attached…	Not Spam

- Transfer learning enabled advancements in NLP
- Use of labelled data to perform fine-tuning

Language ‘Understanding’ within NLP

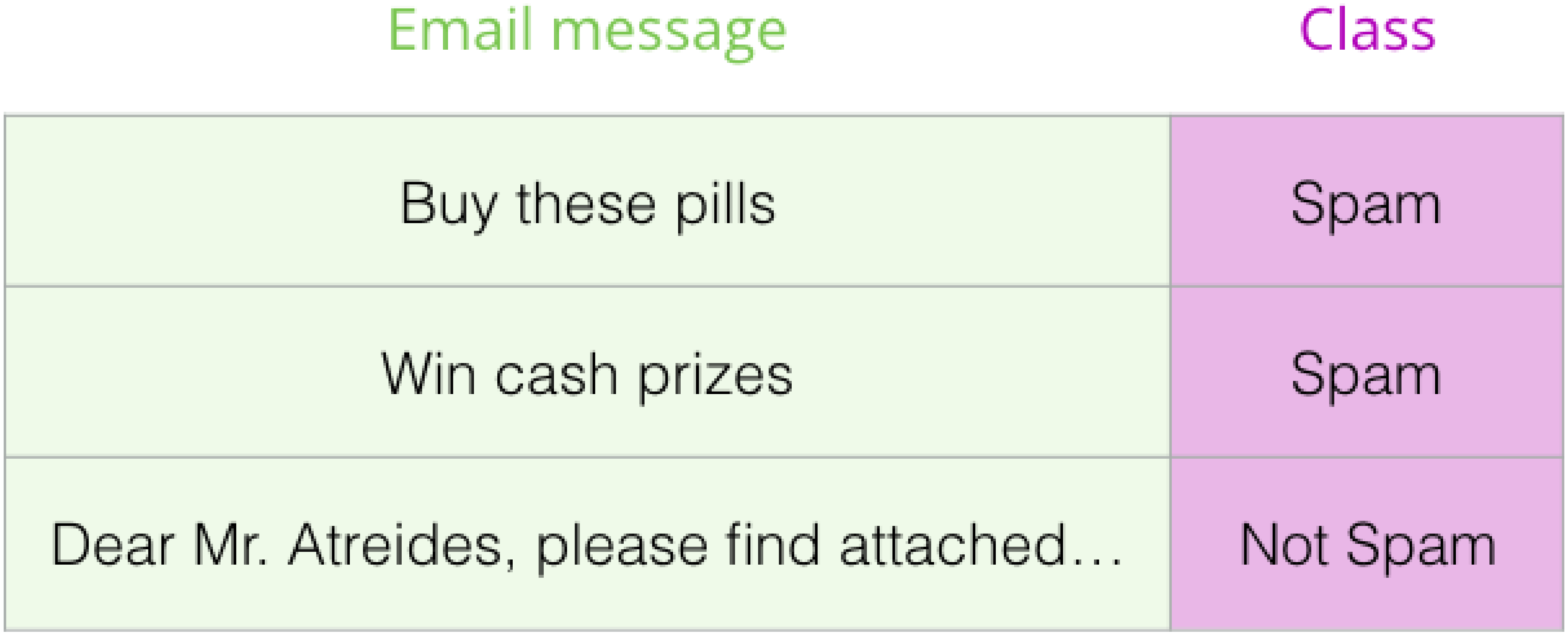
Improved performance in Information Extraction, Retrieval, & Classification

- Entity Recognition - Identify entities, given a piece of text (name, location, abbreviation, longform)
- Retrieval – given a search query, finding products on an e-commerce platform (galaxy s24)
- Classification
  - Identification of Spam, Sentiment, Emotion, Sarcasm, Aggression, Toxicity **[Social NLP]**
- Evaluation
  - Estimation of translation quality, given source sentence and output of Machine Translation (MT) Output
    - Source: I need some coffee.
    - MT Output: ‘मुझे कुछ कॉफ़ी चाहिए’ (*mujhe kuchh coffee chahiye*)
    - Quality: 65

Language ‘Understanding’ within NLP

Improved performance in Information Extraction, Retrieval, &

C



Email message	Class
Buy these pills	Spam
Win cash prizes	Spam
Dear Mr. Atreides, please find attached…	Not Spam

Quality: 65

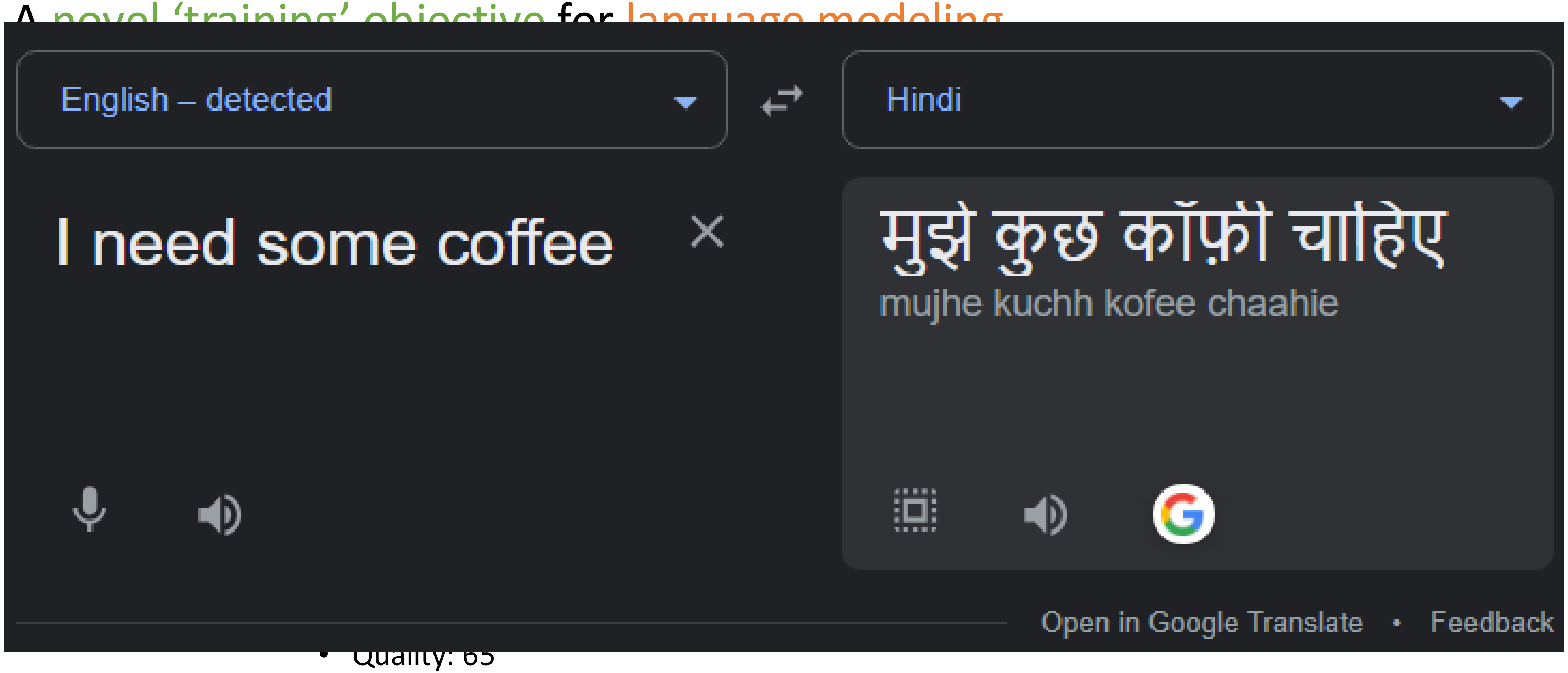
Language ‘Understanding’ within NLP

Improved performance in Information Extraction, Retrieval, & Classification

- Entity Recognition - Identify entities, given a piece of text (name, location, abbreviation, longform)
- Retrieval – given a search query, finding products on an e-commerce platform (galaxy s24)
- Classification
  - Identification of Sentiment, Emotion, Sarcasm, Aggression, Toxicity **[Social NLP]**
- Evaluation
  - Estimation of translation quality, given source sentence and output of Machine Translation (MT) Output
    - Source: I need some coffee.
    - MT Output: ‘मुझे कुछ कॉफ़ी चाहिए’ (*mujhe kuchh coffee chahiye*)
    - Quality: 65

Language ‘Understanding’ within NLP

- A novel ‘training’ objective for language modeling


English – detected

Hindi

I need some coffee

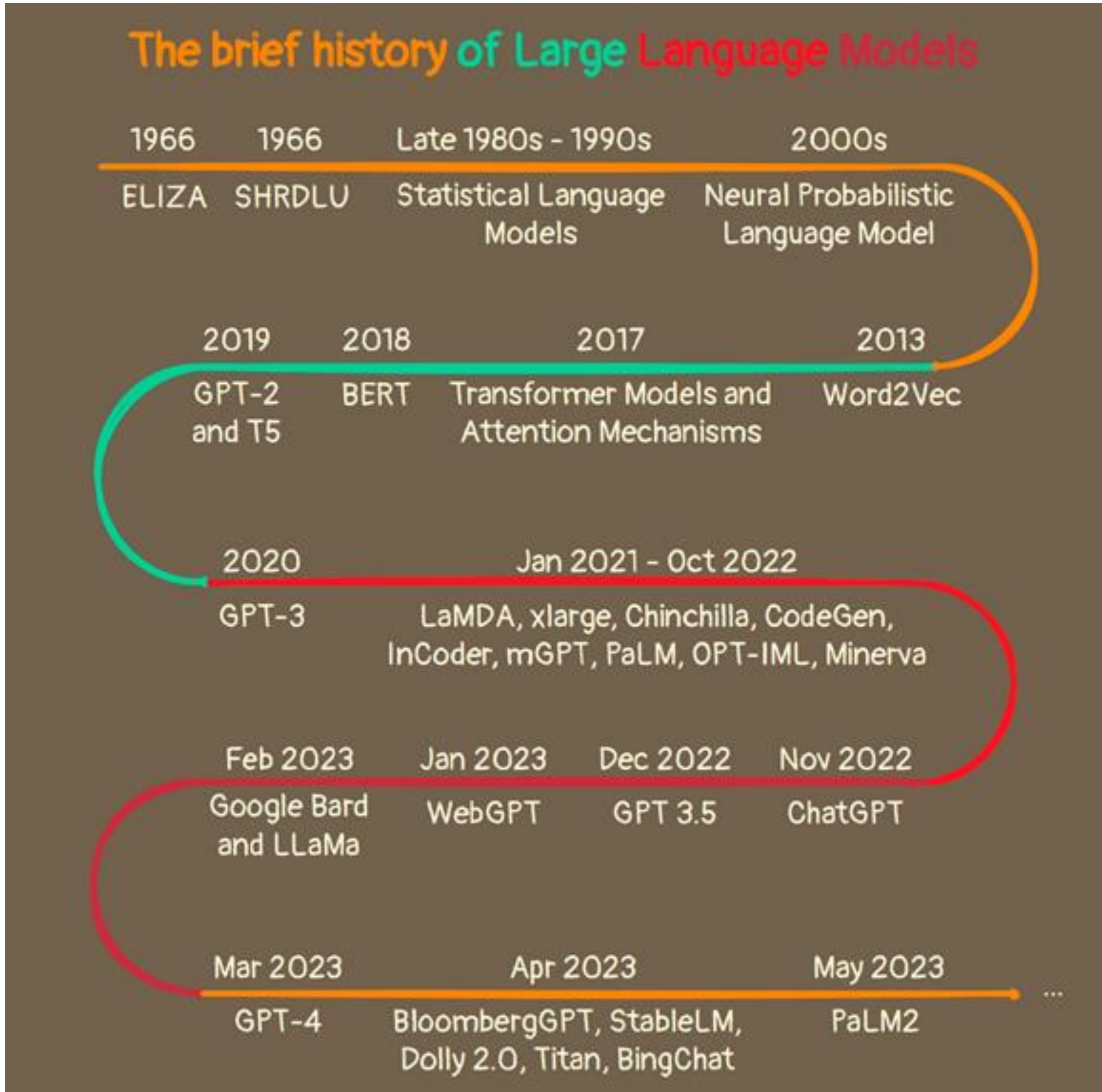
मुझे कुछ कॉफ़ी चाहिए

mujhe kuchh kofee chaahie

    - Quality: 65

Visualizing Language Modeling Evolution

- From 2017 to 2024, 100+ new language models engineered to different,
  - **Sizes** (100 million - 170+ Billion parameters)
  - **Domains** (Healthcare, Finance, Biomedical, …)
  - **Language-specific** (Hindi – *Airavata, OdiaGPT, …*)
  - **Task-oriented** (Dialogue, Translation, eCommerce product relevance,…)



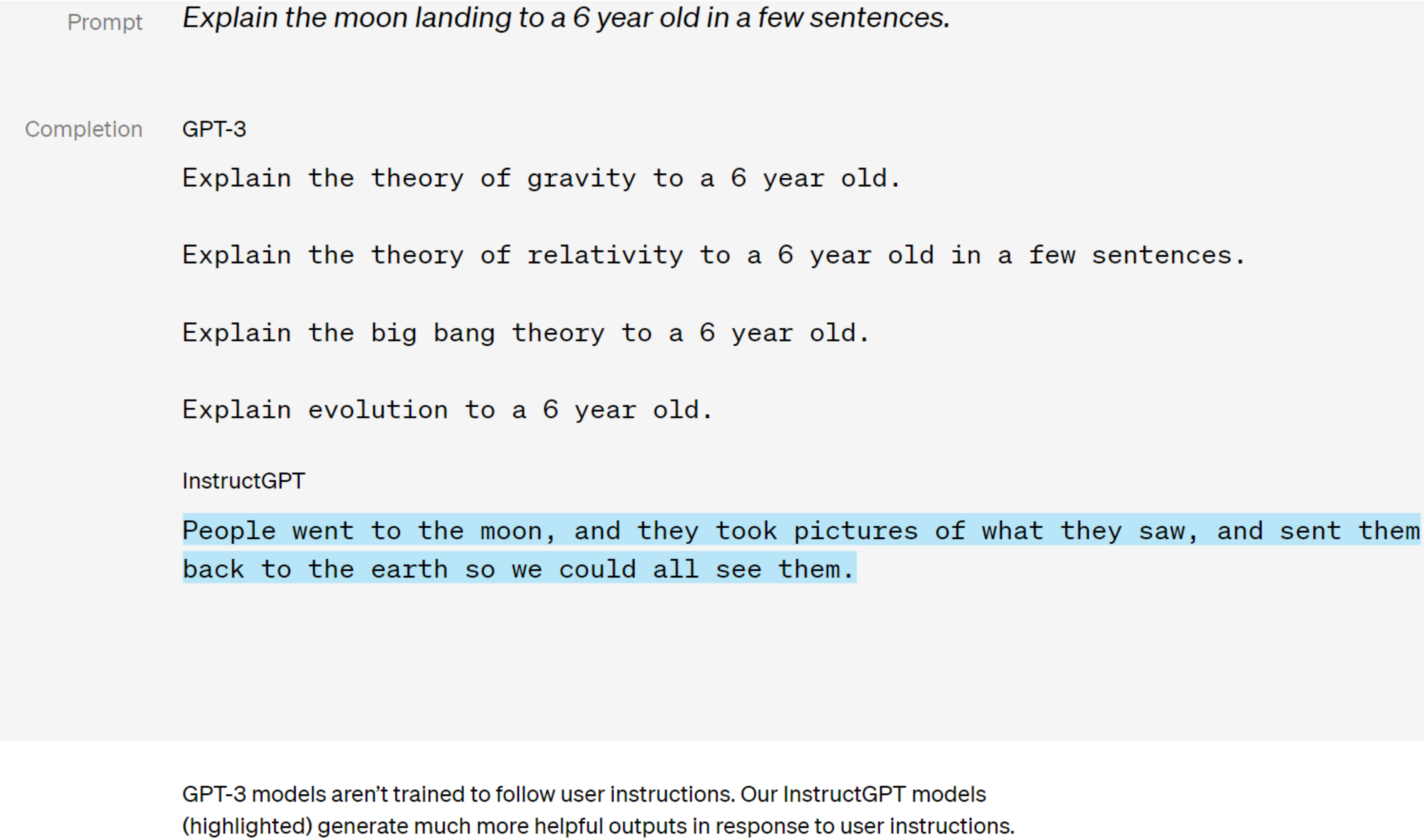
The brief history of Large Language Models

1966 ELIZA
1966 SHRDLU
Late 1980s - 1990s Statistical Language Models
2000s Neural Probabilistic Language Model
2013 Word2Vec
2017 Transformer Models and Attention Mechanisms
2018 BERT
2019 GPT-2 and T5
2020 GPT-3
Jan 2021 - Oct 2022 LaMDA, xlarge, Chinchilla, CodeGen, InCoder, mGPT, PaLM, OPT-IML, Minerva
Nov 2022 ChatGPT
Dec 2022 GPT 3.5
Jan 2023 WebGPT
Feb 2023 Google Bard and LLaMa
Mar 2023 GPT-4
Apr 2023 BloombergGPT, StableLM, Dolly 2.0, Titan, BingChat
May 2023 PaLM2
...

Language Generation within NLP

- Machine Translation, Text Summarization, or Dialogue Generation, …
- **Synthetic Data** Generation
  - Given an example of hate speech on social media, generate a counter narrative
- Data **Labelling**
  - Given a piece of text, provide a label the for text
    - Positive, negative, neutral, sarcastic, non-sarcastic, hateful, non-hateful
- Generating **creative text** like poetry

InstructGPT (steppingstone to ChatGPT)



Prompt *Explain the moon landing to a 6 year old in a few sentences.*

Completion GPT-3

```
Explain the theory of gravity to a 6 year old.

Explain the theory of relativity to a 6 year old in a few sentences.

Explain the big bang theory to a 6 year old.

Explain evolution to a 6 year old.
```

InstructGPT

```
People went to the moon, and they took pictures of what they saw, and sent them
back to the earth so we could all see them.
```

GPT-3 models aren't trained to follow user instructions. Our InstructGPT models (highlighted) generate much more helpful outputs in response to user instructions.

ChatGPT's Penetration: Research & Layman

- GPT-2 and GPT-3's early research deployments
  - State-of-the-art performance on many benchmark datasets.
- GPT-3.5-Turbo to ChatGPT
  - Fine-tuned on instructions like "Translate this sentence for me…"
- Engineering challenges in scaling
  - OpenAI faced challenges in deployment in early stages.
  - Collaboration with Microsoft led to deployment of 'Copilots'.
- Accessibility and impact on non-experts

Publicly available **large language models** grow in **size** and in **number**

- How do we adapt given infrastructure?

Synergy of Engineering and Research

- Low rank adaption
- Parameter efficient fine tuning
- Fusion approaches
- Retrieval Augmented Generation

Language Modeling: Scalable to Infrastructure

Low-rank Adaptation & Parameter efficient fine tuning

➔Language Models are matrices

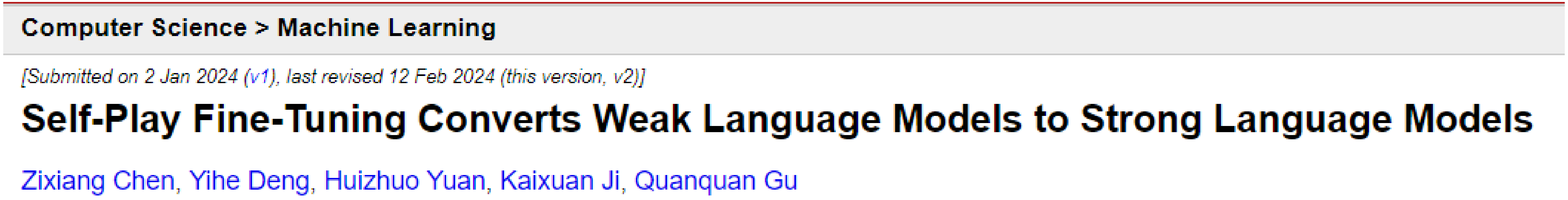
H: Lower rank matrices may contain sufficient information for task/domain

➔Downgrade the matrix built inside the language model

➔Fine tune ‘significantly lower number of parameters’ to obtain **an adapter**.

➔Reportedly up to 95% efficiency, compared to a model 10 times in size.

Efficient Language Modeling



Computer Science > Machine Learning

[Submitted on 2 Jan 2024 (v1), last revised 12 Feb 2024 (this version, v2)]

Self-Play Fine-Tuning Converts Weak Language Models to Strong Language Models

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, Quanquan Gu

SPIN, based on a self-play mechanism, generates its own training data and refines its language generation policy by discerning its own output against human annotated data.

The model trains itself to perform better at various NLP tasks.

Fusion of Language Models

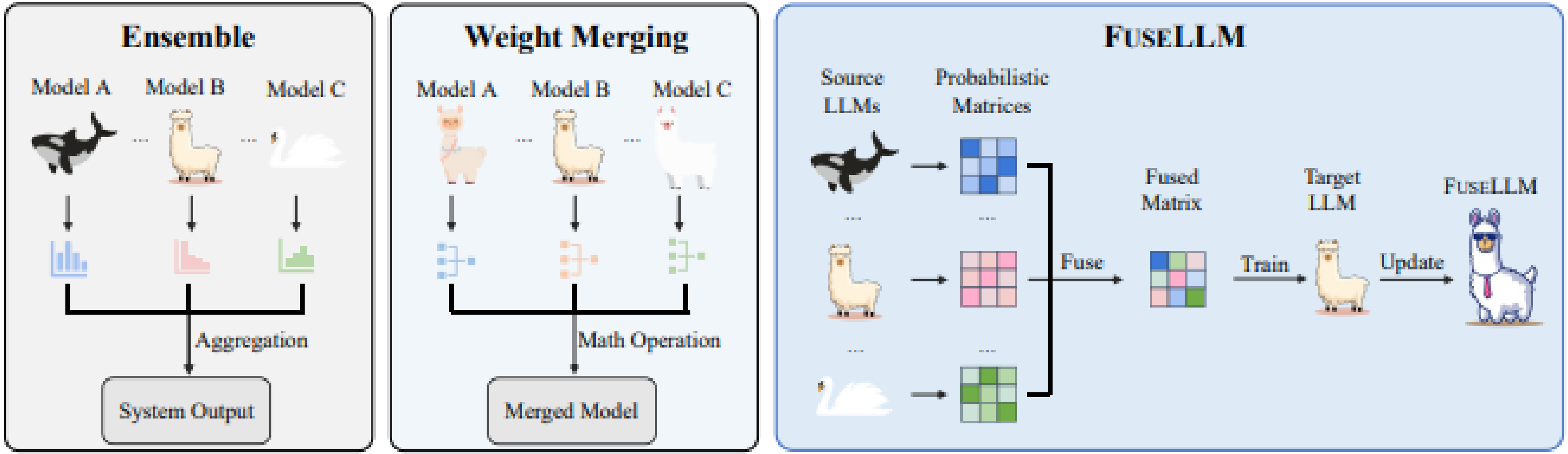
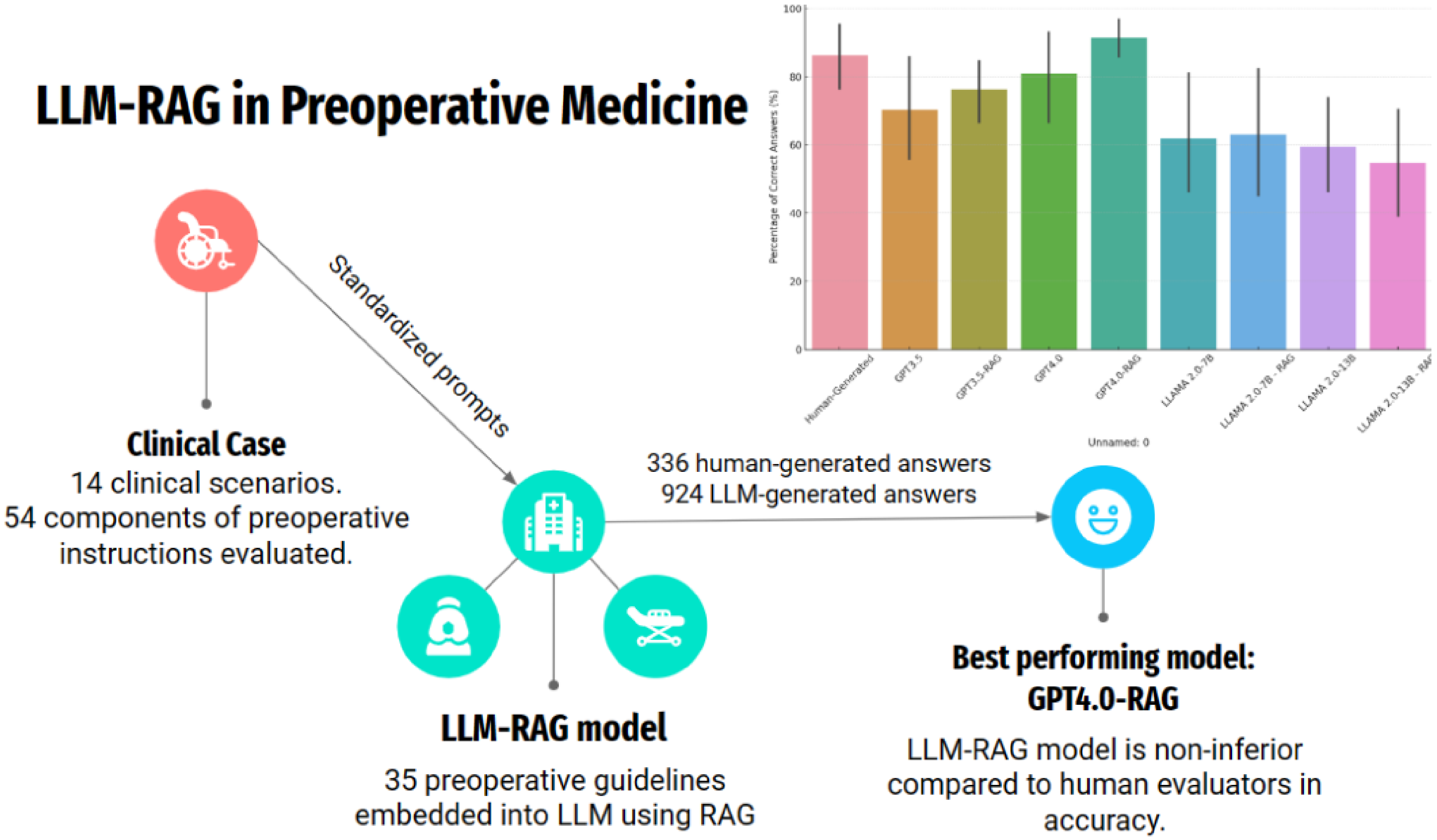


Figure 1: Illustration of conventional model fusion techniques (ensemble and weight merging) and our knowledge fusion approach for LLMs (FuseLLM). Different animal icons represent different LLMs, with various species denoting LLMs possessing differing architectures. FuseLLM externalizes the knowledge from multiple LLMs and transfers their capabilities to a target LLM.

Applications Across Domains - Healthcare

Development and Testing of Retrieval Augmented Generation in Large Language Models -- A Case Study Report

YuHe Ke, Liyuan Jin, Kabilan Elangovan, Hairil Rizal Abdullah, Nan Liu, Alex Tiong Heng Sia, Chai Rick Soh, Joshua Yi Min Tung, Jasmine Chiat Ling Ong, Daniel Shu Wei Ting



LLM-RAG in Preoperative Medicine

Clinical Case
14 clinical scenarios.
54 components of preoperative instructions evaluated.

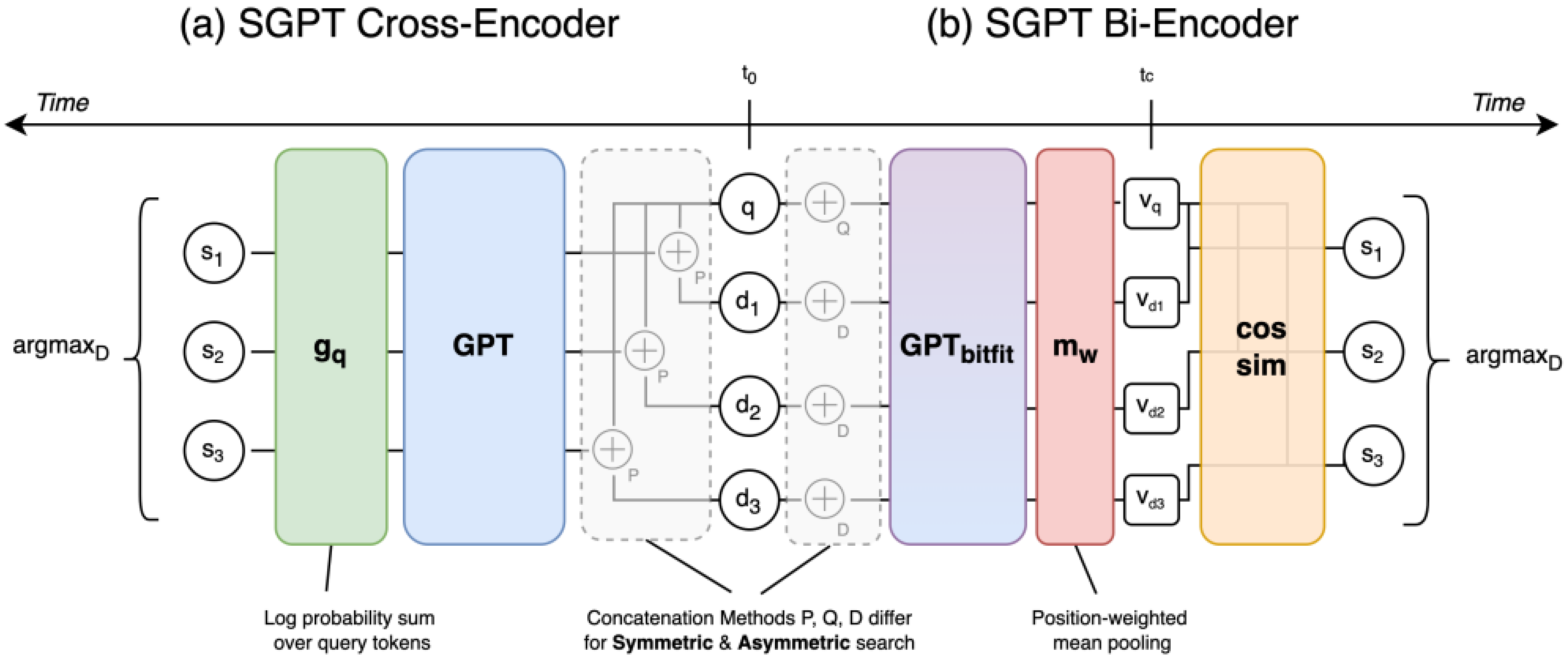
Standardized prompts

LLM-RAG model
35 preoperative guidelines embedded into LLM using RAG

336 human-generated answers
924 LLM-generated answers

Best performing model:
GPT4.0-RAG
LLM-RAG model is non-inferior compared to human evaluators in accuracy.

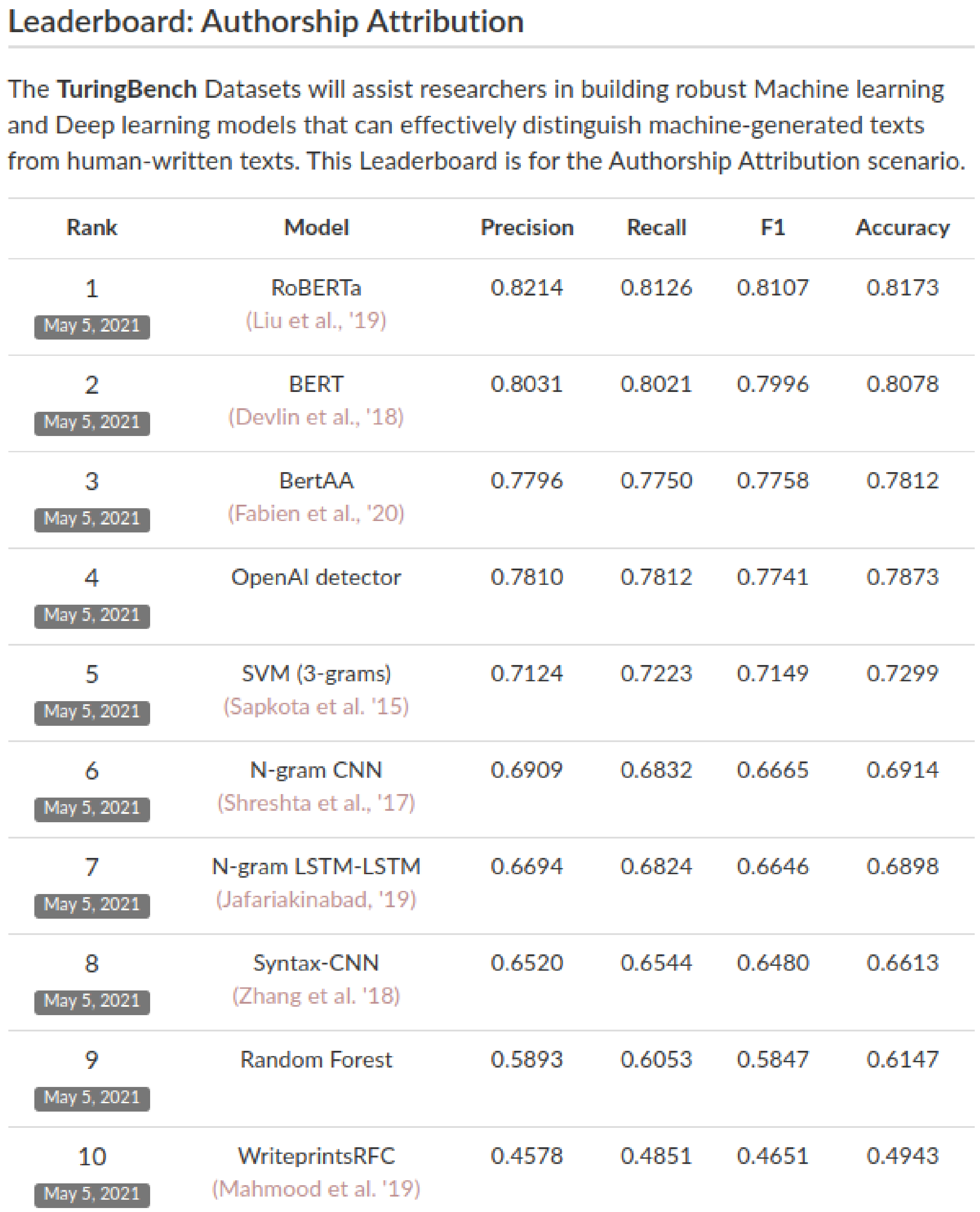
Applications Across Domains - eCommerce



Challenges, Risks and a People-Centred perspective

Challenges for Academia

- Coping with AI-generated text
- Authorship Attribution for Neural Text Generation

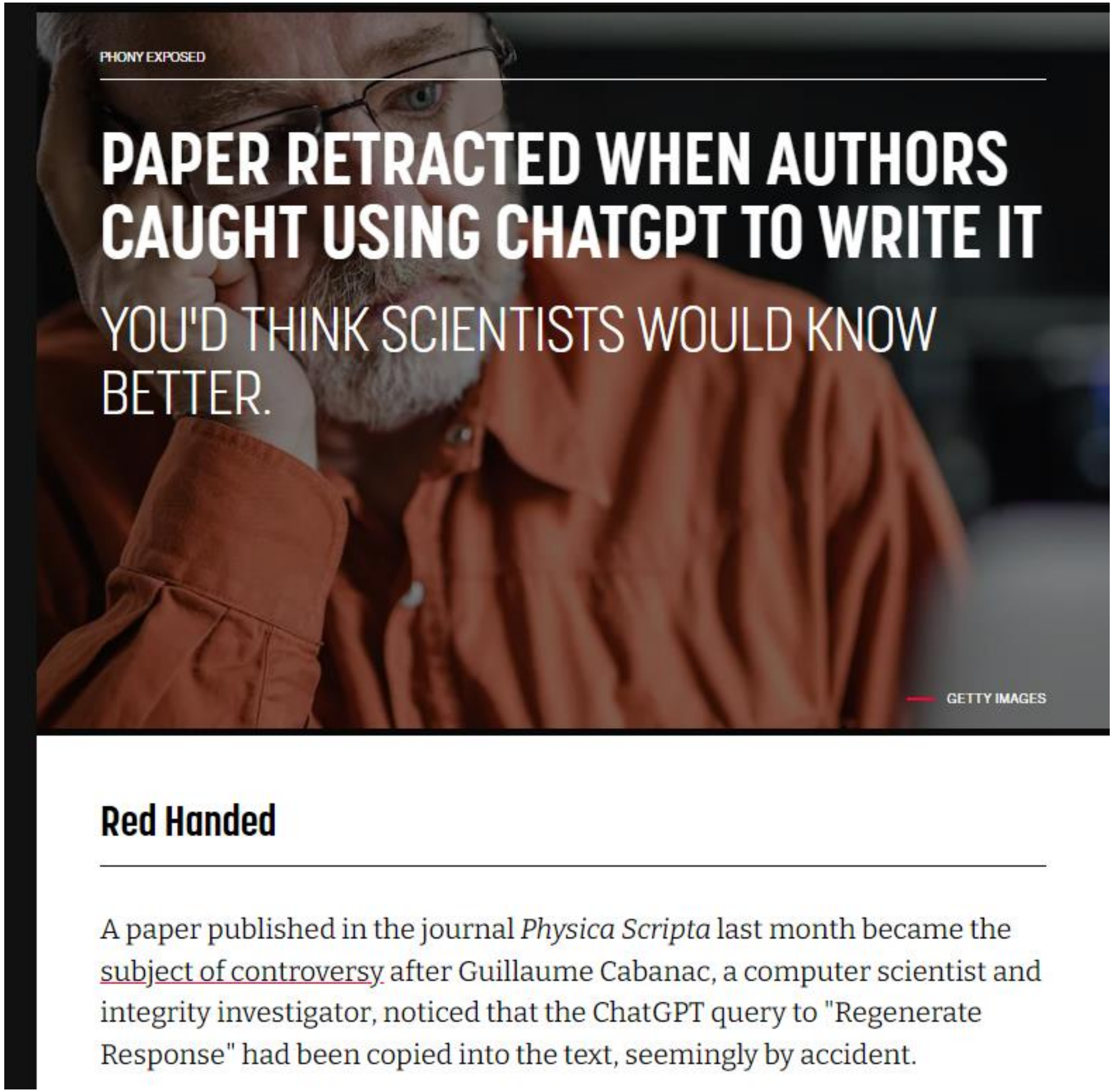


Leaderboard: Authorship Attribution

The **TuringBench** Datasets will assist researchers in building robust Machine learning and Deep learning models that can effectively distinguish machine-generated texts from human-written texts. This Leaderboard is for the Authorship Attribution scenario.

Rank	Model	Precision	Recall	F1	Accuracy
1 May 5, 2021	RoBERTa (Liu et al., '19)	0.8214	0.8126	0.8107	0.8173
2 May 5, 2021	BERT (Devlin et al., '18)	0.8031	0.8021	0.7996	0.8078
3 May 5, 2021	BertAA (Fabien et al., '20)	0.7796	0.7750	0.7758	0.7812
4 May 5, 2021	OpenAI detector	0.7810	0.7812	0.7741	0.7873
5 May 5, 2021	SVM (3-grams) (Sapkota et al. '15)	0.7124	0.7223	0.7149	0.7299
6 May 5, 2021	N-gram CNN (Shreshta et al., '17)	0.6909	0.6832	0.6665	0.6914
7 May 5, 2021	N-gram LSTM-LSTM (Jafariakinabad, '19)	0.6694	0.6824	0.6646	0.6898
8 May 5, 2021	Syntax-CNN (Zhang et al. '18)	0.6520	0.6544	0.6480	0.6613
9 May 5, 2021	Random Forest	0.5893	0.6053	0.5847	0.6147
10 May 5, 2021	WriteprintsRFC (Mahmood et al. '19)	0.4578	0.4851	0.4651	0.4943

Misuse of AI



PHONY EXPOSED

PAPER RETRACTED WHEN AUTHORS CAUGHT USING CHATGPT TO WRITE IT

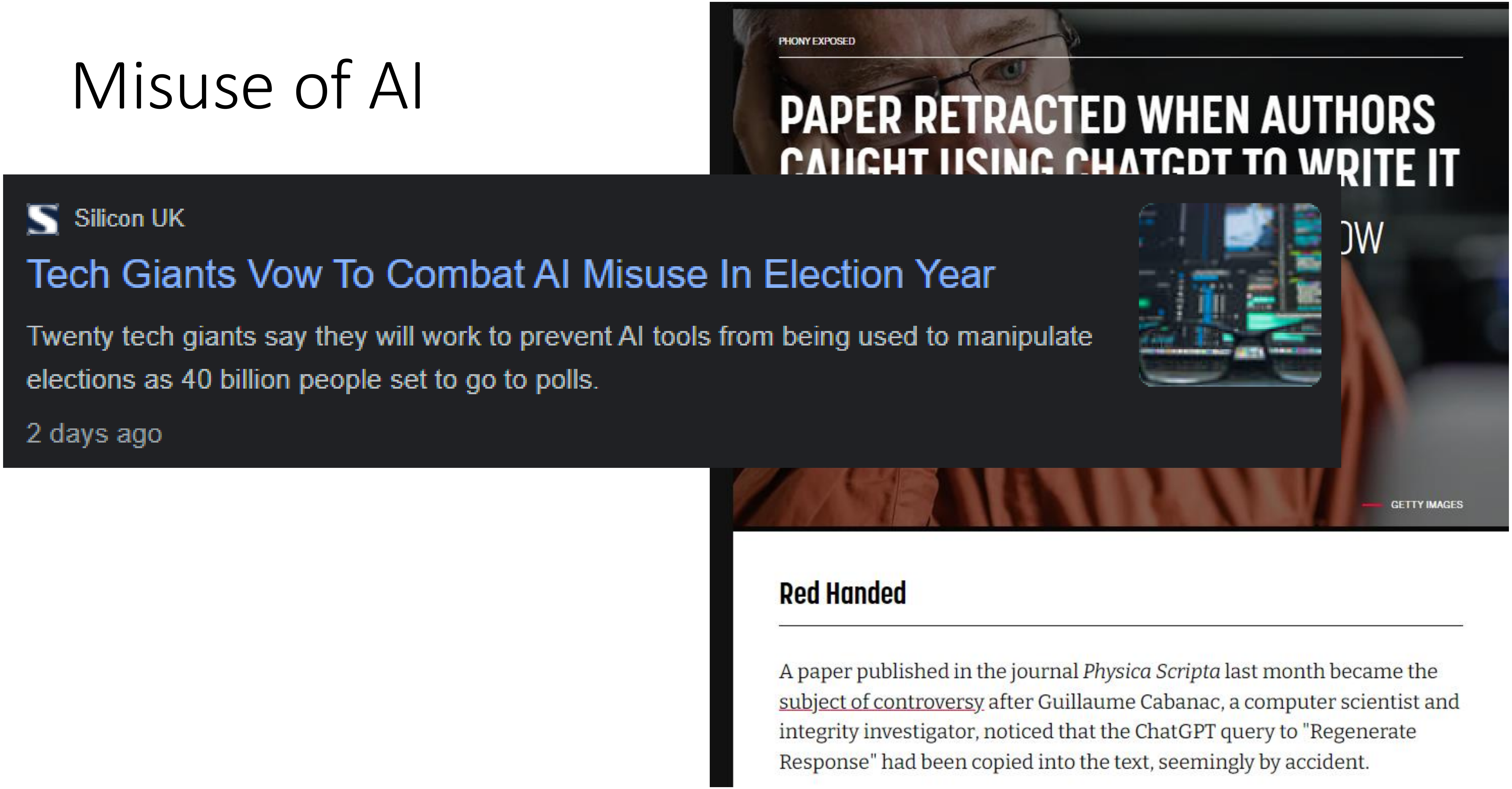
YOU'D THINK SCIENTISTS WOULD KNOW BETTER.

GETTY IMAGES

Red Handed

A paper published in the journal *Physica Scripta* last month became the subject of controversy after Guillaume Cabanac, a computer scientist and integrity investigator, noticed that the ChatGPT query to "Regenerate Response" had been copied into the text, seemingly by accident.

Misuse of AI



Silicon UK

Tech Giants Vow To Combat AI Misuse In Election Year

Twenty tech giants say they will work to prevent AI tools from being used to manipulate elections as 40 billion people set to go to polls.

2 days ago

PHONY EXPOSED

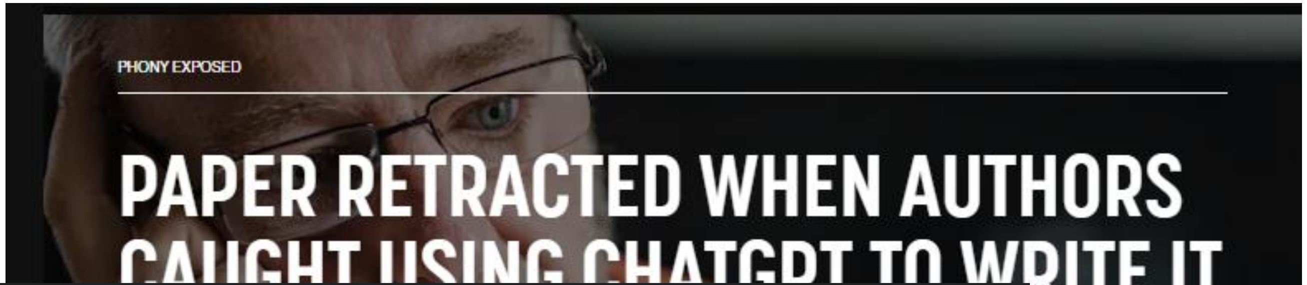
PAPER RETRACTED WHEN AUTHORS CAUGHT USING CHATGPT TO WRITE IT

GETTY IMAGES

Red Handed

A paper published in the journal *Physica Scripta* last month became the subject of controversy after Guillaume Cabanac, a computer scientist and integrity investigator, noticed that the ChatGPT query to "Regenerate Response" had been copied into the text, seemingly by accident.

Misuse of AI



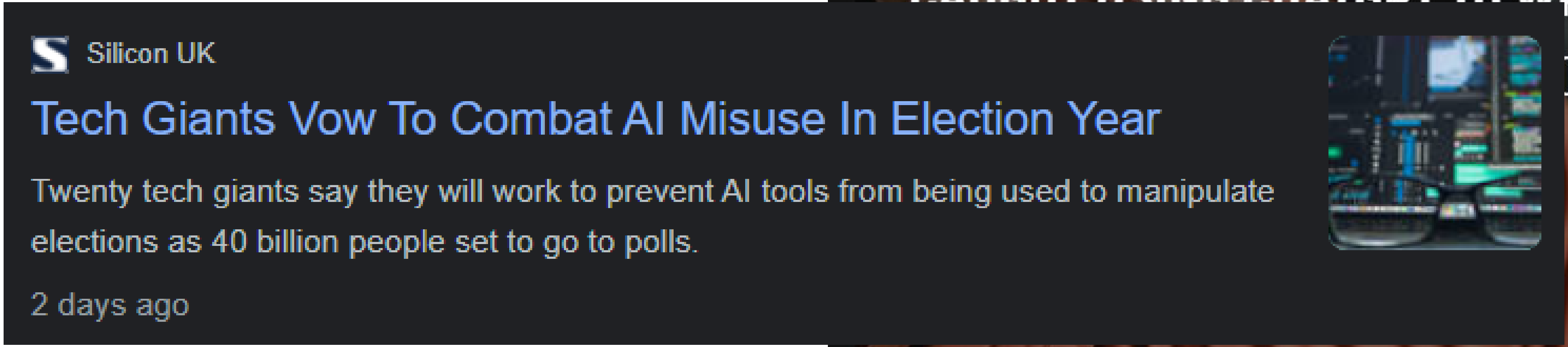
PHONY EXPOSED

PAPER RETRACTED WHEN AUTHORS CAUGHT USING CHATGPT TO WRITE IT

OW

GETTY IMAGES

ecame the
scientist and
enerate
ent.

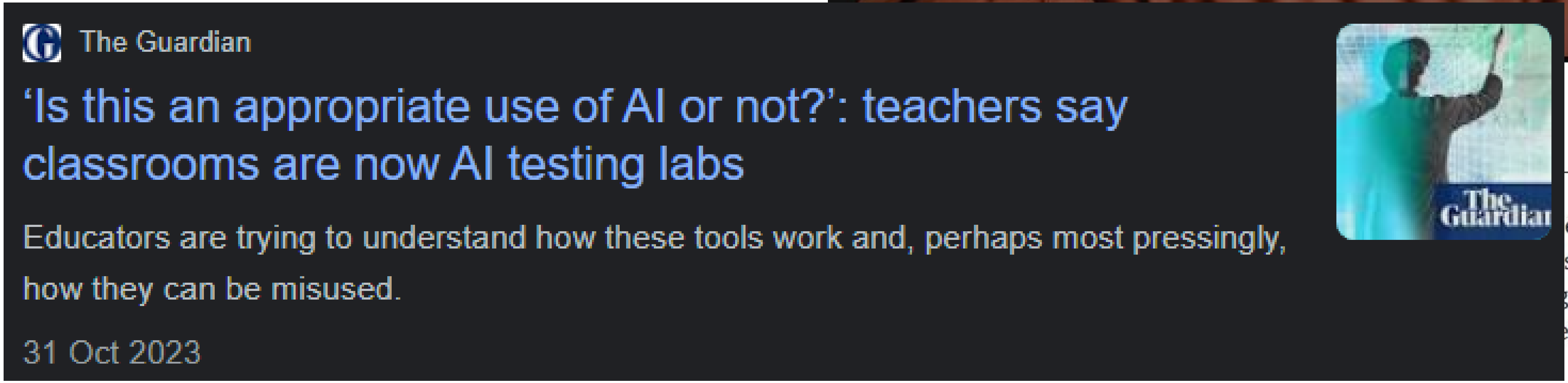


Silicon UK

Tech Giants Vow To Combat AI Misuse In Election Year

Twenty tech giants say they will work to prevent AI tools from being used to manipulate elections as 40 billion people set to go to polls.

2 days ago



The Guardian

‘Is this an appropriate use of AI or not?’: teachers say classrooms are now AI testing labs

Educators are trying to understand how these tools work and, perhaps most pressingly, how they can be misused.

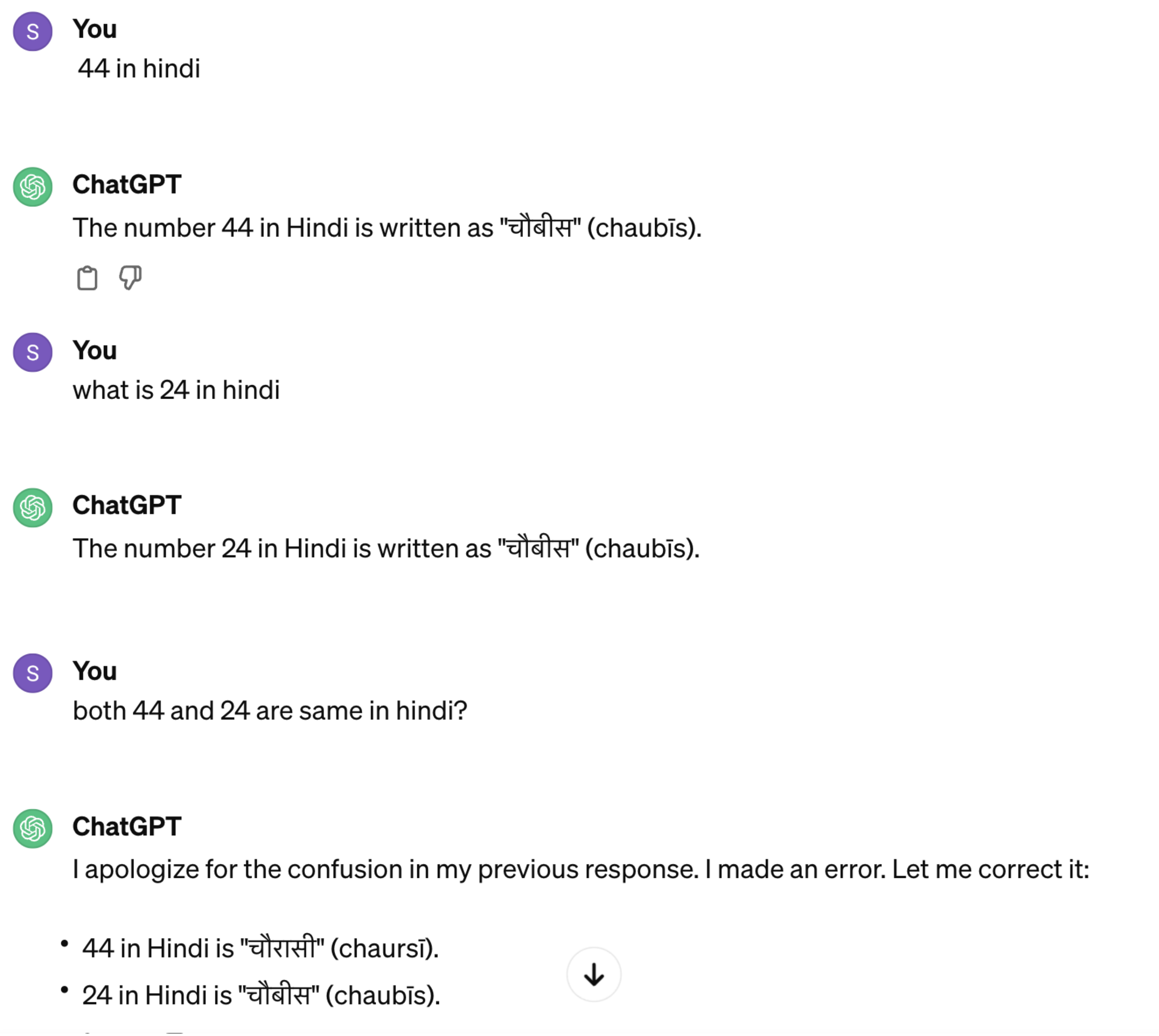
31 Oct 2023

Risks of Generative AI

- Data curation – Primary issue
  - Training at scale requires large amounts of clean data.
  - Data for training language models comprises of anything on the web – from Wikipedia articles to Reddit posts to GitHub code.
- Deepfakes- Textual and Video
- Generation of Hateful and Toxic content

- Safety and Content Moderation
  - Safety evaluation

  - LLM Alignment



You

44 in hindi

ChatGPT

The number 44 in Hindi is written as "चौबीस" (chaubīs).

You

what is 24 in hindi

ChatGPT

The number 24 in Hindi is written as "चौबीस" (chaubīs).

You

both 44 and 24 are same in hindi?

ChatGPT

I apologize for the confusion in my previous response. I made an error. Let me correct it:

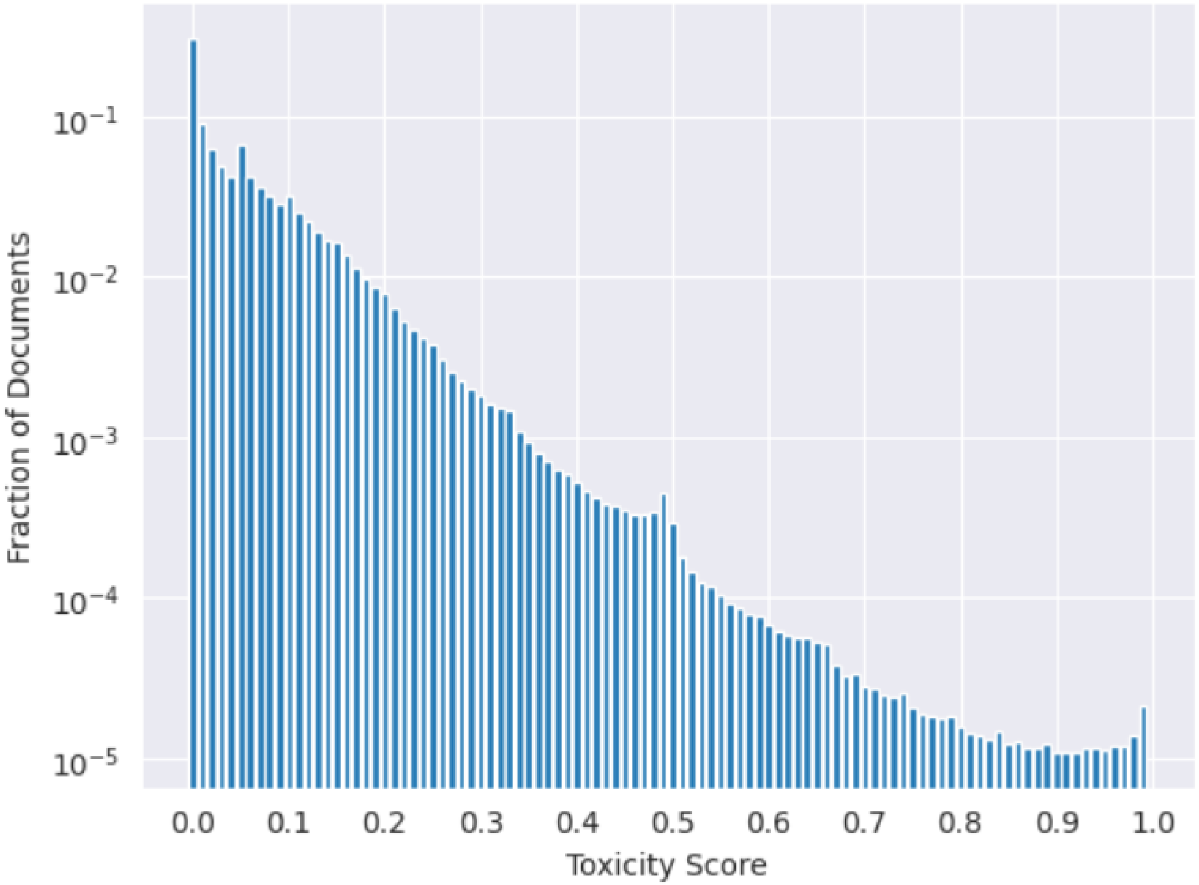
- 44 in Hindi is "चौरासी" (chaursī).
- 24 in Hindi is "चौबीस" (chaubīs).

; on the web – from

Example: Llama-v2 Data

Pretraining data toxicity

"To allow for better downstream generalization, we chose not to scrub toxic data from pretraining"



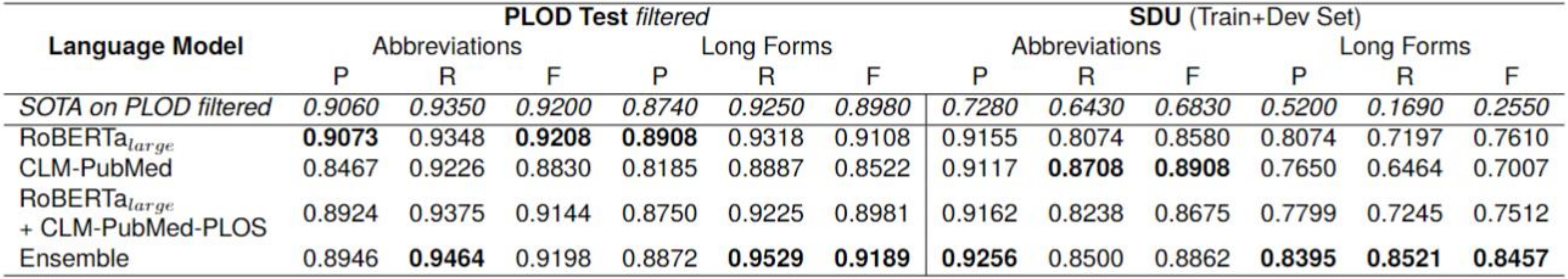
People-Centred AI Initiatives

- Low resource and endangered languages.
- Efficient and Scalable Language Modeling
- Large Language Model evaluation for bias, toxicity, and misinformation.
- Detection of Aggression and Offense on social media, including political aggression

On Efficient Information Extraction

Character level language models, (almost) as good as Transformer encoders.

Cost effective solution (computational, storage, time)

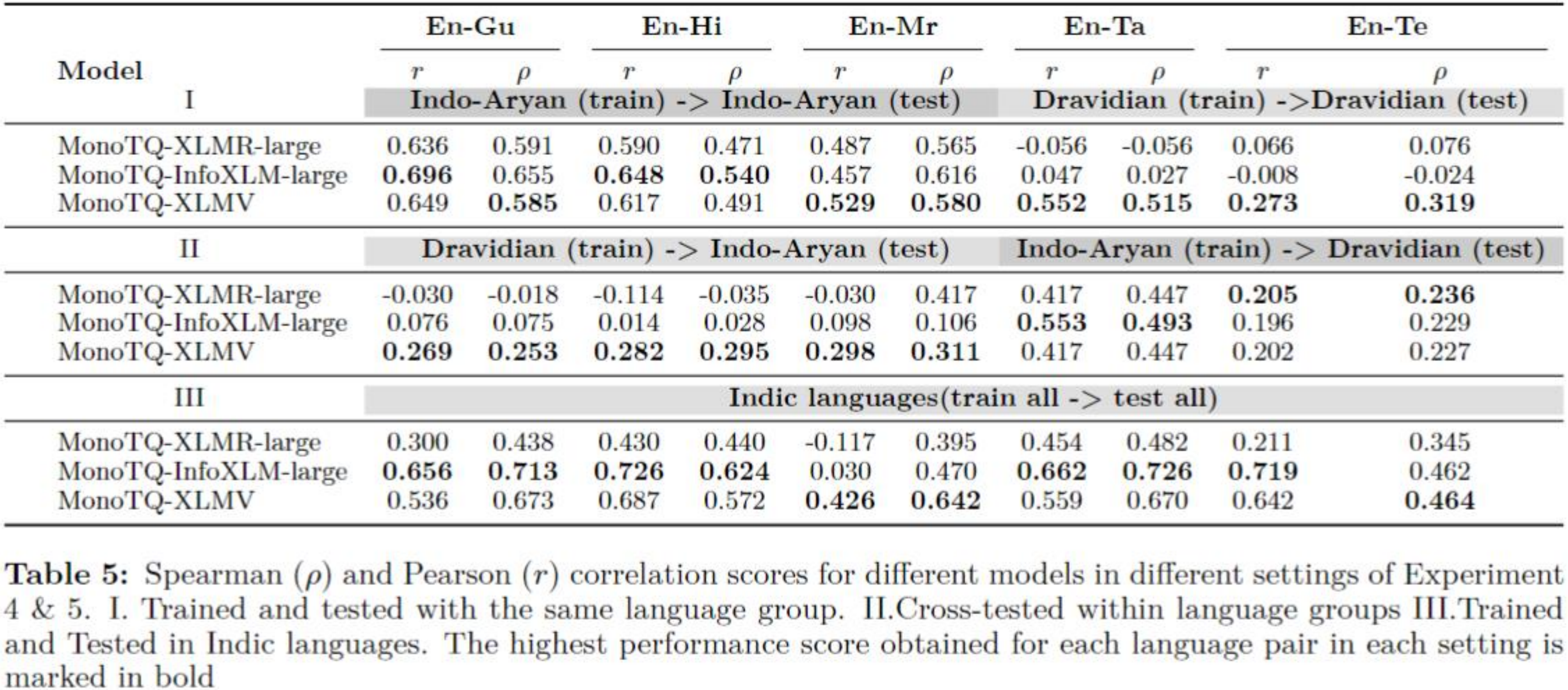


	PLOD Test *filtered*						**SDU** (Train+Dev Set)					
Language Model	Abbreviations			Long Forms			Abbreviations			Long Forms		
	P	R	F	P	R	F	P	R	F	P	R	F
SOTA on PLOD filtered	*0.9060*	*0.9350*	*0.9200*	*0.8740*	*0.9250*	*0.8980*	*0.7280*	*0.6430*	*0.6830*	*0.5200*	*0.1690*	*0.2550*
RoBERTa$_{large}$	**0.9073**	0.9348	**0.9208**	**0.8908**	0.9318	0.9108	0.9155	0.8074	0.8580	0.8074	0.7197	0.7610
CLM-PubMed	0.8467	0.9226	0.8830	0.8185	0.8887	0.8522	0.9117	**0.8708**	**0.8908**	0.7650	0.6464	0.7007
RoBERTa$_{large}$ + CLM-PubMed-PLOS	0.8924	0.9375	0.9144	0.8750	0.9225	0.8981	0.9162	0.8238	0.8675	0.7799	0.7245	0.7512
Ensemble	0.8946	**0.9464**	0.9198	0.8872	**0.9529**	**0.9189**	**0.9256**	0.8500	0.8862	**0.8395**	**0.8521**	**0.8457**

Table 5: Abbreviation Detection performance using various language model combinations evaluated using Precision (P), Recall (R), and F1-score (F), **trained on re-annotated, *filtered* PLOD v2 BIO data**, and tested on both.

On Efficient Quality Estimation

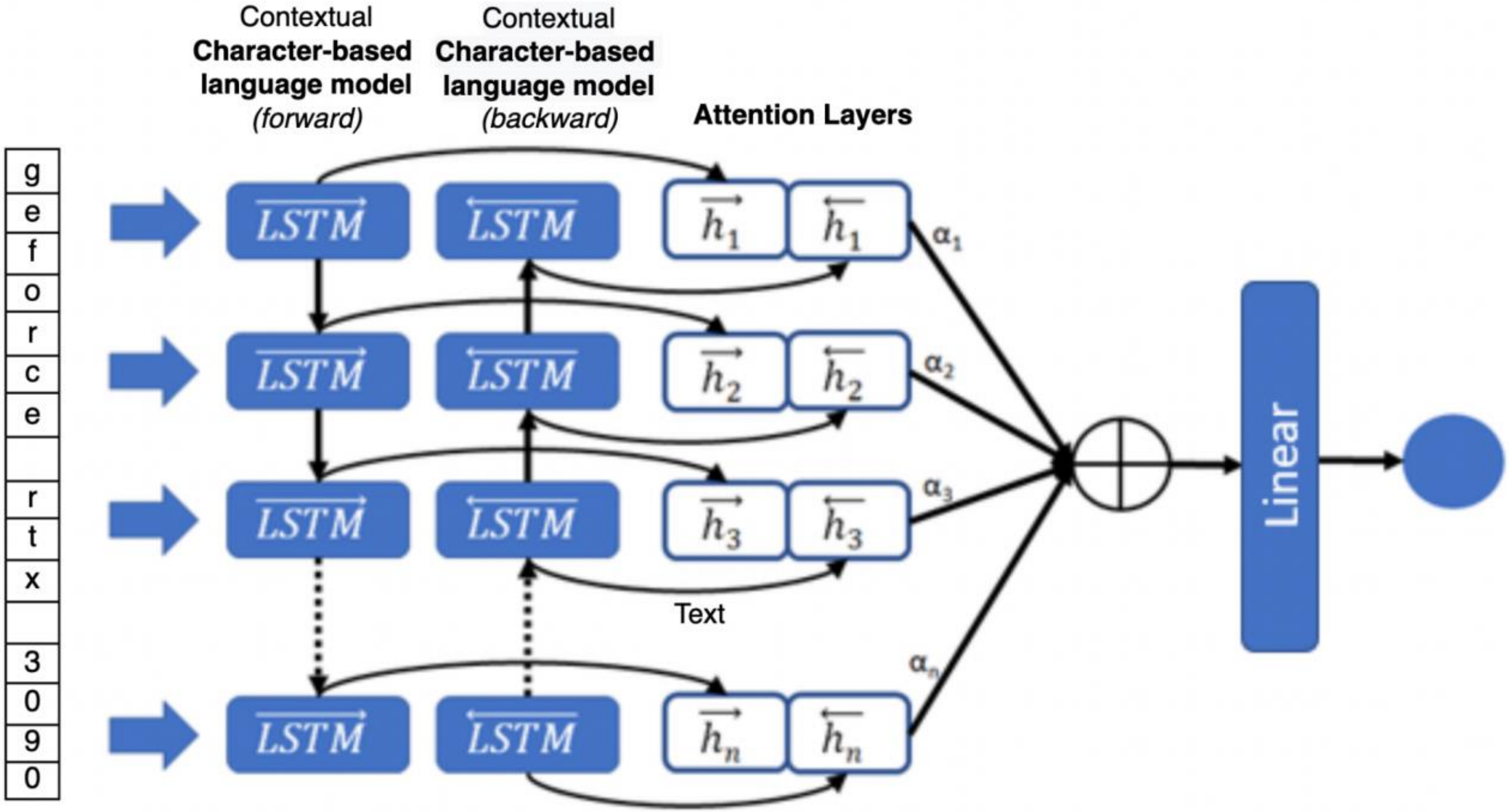
- ‘Language family’ dependent models for evaluation of MT models.



Model	En-Gu		En-Hi		En-Mr		En-Ta		En-Te	
	r	ρ	r	ρ	r	ρ	r	ρ	r	ρ
I	**Indo-Aryan (train) -> Indo-Aryan (test)**						**Dravidian (train) ->Dravidian (test)**			
MonoTQ-XLMR-large	0.636	0.591	0.590	0.471	0.487	0.565	-0.056	-0.056	0.066	0.076
MonoTQ-InfoXLM-large	**0.696**	0.655	**0.648**	**0.540**	0.457	0.616	0.047	0.027	-0.008	-0.024
MonoTQ-XLMV	0.649	**0.585**	0.617	0.491	**0.529**	**0.580**	**0.552**	**0.515**	**0.273**	**0.319**
II	**Dravidian (train) -> Indo-Aryan (test)**						**Indo-Aryan (train) -> Dravidian (test)**			
MonoTQ-XLMR-large	-0.030	-0.018	-0.114	-0.035	-0.030	0.417	0.417	0.447	**0.205**	**0.236**
MonoTQ-InfoXLM-large	0.076	0.075	0.014	0.028	0.098	0.106	**0.553**	**0.493**	0.196	0.229
MonoTQ-XLMV	**0.269**	**0.253**	**0.282**	**0.295**	**0.298**	**0.311**	0.417	0.447	0.202	0.227
III	**Indic languages(train all -> test all)**									
MonoTQ-XLMR-large	0.300	0.438	0.430	0.440	-0.117	0.395	0.454	0.482	0.211	0.345
MonoTQ-InfoXLM-large	**0.656**	**0.713**	**0.726**	**0.624**	0.030	0.470	**0.662**	**0.726**	**0.719**	0.462
MonoTQ-XLMV	0.536	0.673	0.687	0.572	**0.426**	**0.642**	0.559	0.670	0.642	**0.464**

Table 5: Spearman (ρ) and Pearson (r) correlation scores for different models in different settings of Experiment 4 & 5. I. Trained and tested with the same language group. II.Cross-tested within language groups III.Trained and Tested in Indic languages. The highest performance score obtained for each language pair in each setting is marked in bold

Efficient Retrieval and Ranking (eBay Inc.)



Prioritizing Toxicity and Truthfulness

- Non-toxic Data for pre-training

- Generation of verifiable factual information – trustworthiness in AI

- Non-toxic generation (even if prompted to do so)

AIRAVATA: INTRODUCING HINDI INSTRUCTION-TUNED LLM

Jay Gala[1] **Thanmay Jayakumar**[1] **Jaavid Aktar Husain**[1,3] **Aswanth Kumar**[4]
Mohammed Safi Ur Rahman Khan[1] **Diptesh Kanojia**[5] **Ratish Puduppully**[6]
Mitesh M. Khapra[1,2] **Raj Dabre**[7] **Rudra Murthy**[8] **Anoop Kunchukuttan**[1,2,9]

[1]Nilekani Centre at AI4Bharat [2]IIT Madras [3]IIIT D&M Kancheepuram
[4]Flipkart [5]University of Surrey [6]A*STAR [7]NICT [8]IBM Research [9]Microsoft

`https://ai4bharat.github.io/airavata`

Collaboration with SWLEOC

South West London Elective Orthopaedic Centre (SWLEOC)

Using **Large Language Models** to derive insights towards personalized healthcare

Derive insights from various data sources such as pre-operative assessments, reports, and anonymized patient data.

Closing Thoughts

- Generative AI is as much engineering, as it is research.
- Recent advancements can help engineer solutions to problems across domains
- Generative AI and growth of language models poses new challenges and risks to education, social fabric, and mental health.
- Need for people-centred approach to language modeling
- We are doing our part! ☺

Acknowledgements

- Illustrations from Jay Alammar’s blog on Illustrated BERT, Transformers, and so on.
- Slide on NLP problems confluence – Prof Pushpak Bhattacharyya
- Illustrations from other sources on web too.

Thank you for listening

- Let us discuss!