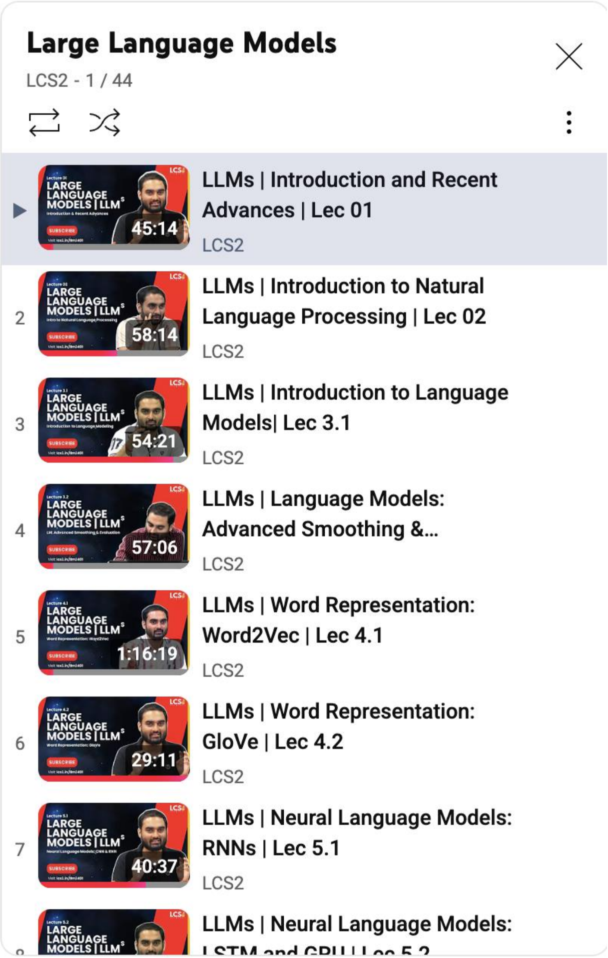
Advances in Large Language Models

ELL8299 · AIL861 · ELL881

Tanmoy Chakraborty
Associate Professor, IIT Delhi
https://tanmoychak.com/

Last year’s offering

https://lcs2.in/llm2401



LLMs | Introduction and Recent Advances | Lec 01

https://www.youtube.com/watch?v=zMn37YxPD6I&list=PLqGkIjcOyrGnjyBHl4GE2S9kX47X96FH-

Course Instructors

Tanmoy Chakraborty
IIT Delhi

Yatin Nandwani
IBM Research

Dinesh Raghu
IBM Research

Sourish Dasgupta
DA-IICT

Gaurav Pandey
IBM Research

Manish Gupta
Microsoft

Course TAs

Shashank Agarwal
PhD student,
IIT Delhi

Aswini Kumar Padhi
PhD student,
IIT Delhi

Prottay Kumar Adhikary
PhD student,
IIT Delhi

Anwoy Chatterjee
PhD student,
IIT Delhi

Course Directives

- Slot **H** (Mon, Wed: 11-12; Thu: 12-13)
- Website: https://lcs2.in/llm2501
- Room: Bharti-301

Marks distribution (tentative)

- Minor: 15%
- Major: 25%
- Quiz (2): 20%
- Assignment (1): 15%
- Mini-project: 25% (group-wise)

- **Audit**: B- (threshold to pass the course)
- **Grading Scheme**: TBD

Course Project

- Some problem statements, and datasets will be floated soon*
- Each group should consist of 1-2 students
- You need to
  - develop models
  - evaluate your models
  - prepare presentation
  - write tech report

Students are encouraged to publish their projects in good conferences/journals

* You are welcome to propose a new idea if you find it fascinating to be qualified for a course project. Instructor opines!

Course Project

- Some problem statements, and datasets will be floated soon*
- Each group should consist of 1-2 students
- You need to
  - develop models
  - evaluate your models
  - prepare presentation
  - write tech report

Students are encouraged to publish their projects in good conferences/journals

Deliverables:

1. Final project report (**10%**), 8 pages ACL format. Encouraged to arXiv
2. Repo of dataset and source code (**5%**)
3. Final project presentation (**10%**)

* You are welcome to propose a new idea if you find it fascinating to be qualified for a course project. Instructor opines!

Do Not Plagiarize !

Academic Integrity is of utmost importance. If anyone is found cheating/plagiarizing, it will result in negative penalty (and possibly even more: an F grade or even DisCo).

Collaborate. But do NOT cheat.

- Assignments to be done individually.
- Do not share any part of code.
- Do not copy any part of report from any online resources or published works.
- If you reuse other's works, always cite.
- If you discuss with others about assignment or outside your group for project, mention their names in the report.
- Do not use GenAI tools (like, ChatGPT).

We will check for pairwise plagiarism in submitted assignment code files among you all.

We will also check the probability of any submitted content being AI generated.

Project reports will be checked for plagiarism across all web resources.

Course Content

- This is an **advanced graduate course** and we will be teaching and discussing state-of-the-art papers and recent advances in the field of large language models.
- **All the students are expected to come to the class regularly.**

Last Year's Course Content

Basics	Architecture	Learnability	User Acceptability	Ethics and Misc.
• Introduction • Intro to NLP • Intro to Language Models (LMs) • Word Embeddings (Word2Vec, GloVE) • Neural LMs (CNN, RNN, Seq2Seq, Attention)	• Intro to Transformer • Decoder-only LM, Prefix LM, Decoding strategies • Encoder-only LM, Encoder-decoder LM • Advanced Attention • Mixture of Experts	• Scaling laws • Instruction fine-tuning • In-context learning • Alignment • Distillation and PEFT • Efficient/Constraint LM inference	• RAG • Tool-augmented LMs • Reasoning • Vision Language Models • Handling long context • Model editing	• Interpretability • Bias and Toxicity

But the state of LLM space has evolved since last year…so we have updated this year’s content

Course Content

Fundamentals

- Course Introduction
- Introduction to Transformers
- Pre-training and Post-training Strategies
- Alignment of Language Models

Course Content

Fundamentals

- Course Introduction
- Introduction to Transformers
- Pre-training and Post-training Strategies
- Alignment of Language Models

Efficiency

- Efficient Design, Training and Inference in LLMs
- Parameter Efficient Fine-Tuning of LLMs
- Model Compression

Course Content

Fundamentals	Efficiency	Augmentation & Reasoning
• Course Introduction • Introduction to Transformers • Pre-training and Post-training Strategies • Alignment of Language Models	• Efficient Design, Training and Inference in LLMs • Parameter Efficient Fine-Tuning of LLMs • Model Compression	• Retrieval-Augmented Language Models • LLM Agents • Large Reasoning Models (LRMs)

Course Content

Fundamentals
- Course Introduction
- Introduction to Transformers
- Pre-training and Post-training Strategies
- Alignment of Language Models

Efficiency
- Efficient Design, Training and Inference in LLMs
- Parameter Efficient Fine-Tuning of LLMs
- Model Compression

Augmentation & Reasoning
- Retrieval-Augmented Language Models
- LLM Agents
- Large Reasoning Models (LRMs)

Alternate Paradigms
- Multimodal Models
- Alternate LLM Architectures

Course Content

Fundamentals

- Course Introduction
- Introduction to Transformers
- Pre-training and Post-training Strategies
- Alignment of Language Models

Efficiency

- Efficient Design, Training and Inference in LLMs
- Parameter Efficient Fine-Tuning of LLMs
- Model Compression

Augmentation & Reasoning

- Retrieval-Augmented Language Models
- LLM Agents
- Large Reasoning Models (LRMs)

Alternate Paradigms

- Multimodal Models
- Alternate LLM Architectures

Miscellaneous

- Physics of Language Models
- Interpretability
- Ethics and Conclusion

Pre-Requisites

- Excitement about language!
- Willingness to learn

Pre-Requisites

- Excitement about language!
- Willingness to learn

Mandatory	Desirable
• Data Structures & Algorithms • Machine Learning • Python programming	• NLP • Deep learning

Pre-Requisites

- Excitement about language!
- Willingness to learn

Mandatory	Desirable
• Data Structures & Algorithms • Machine Learning • Python programming	• NLP • Deep learning

This course will NOT cover:

- Details of NLP (ELL884: https://lcs2.in/nlp2402), Machine Learning and Deep Learning
- Coding practice
- Generative models for modalities other than text

Pre-

You are advised to study the **first 10 lectures (till Lec 6.1)** of the previous year's course playlist before the **next class on August 4**. Otherwise, you will not be able to follow. Here's the link to the playlist:

WE WILL BE COVERING DIFFERENT TOPICS THIS YEAR COMPARED TO PREVIOUS YEAR, AND VIDEOS WILL NOT BE UPLOADED IMMEDIATELY. SO DON'T MISS THE CLASSES!

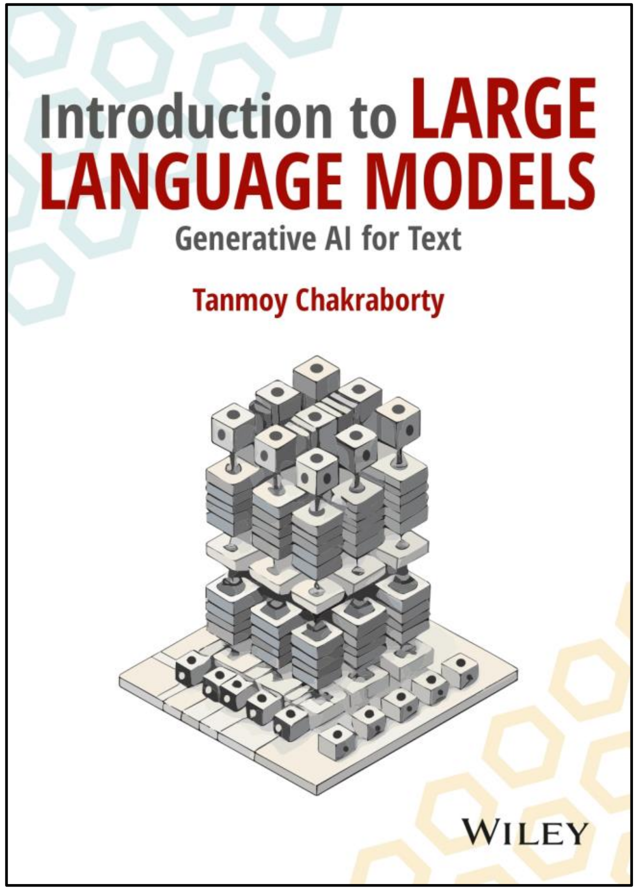
This co

- Details
- Coding
- Genera

Textbook

Introduction to Large Language Models,
Tanmoy Chakraborty

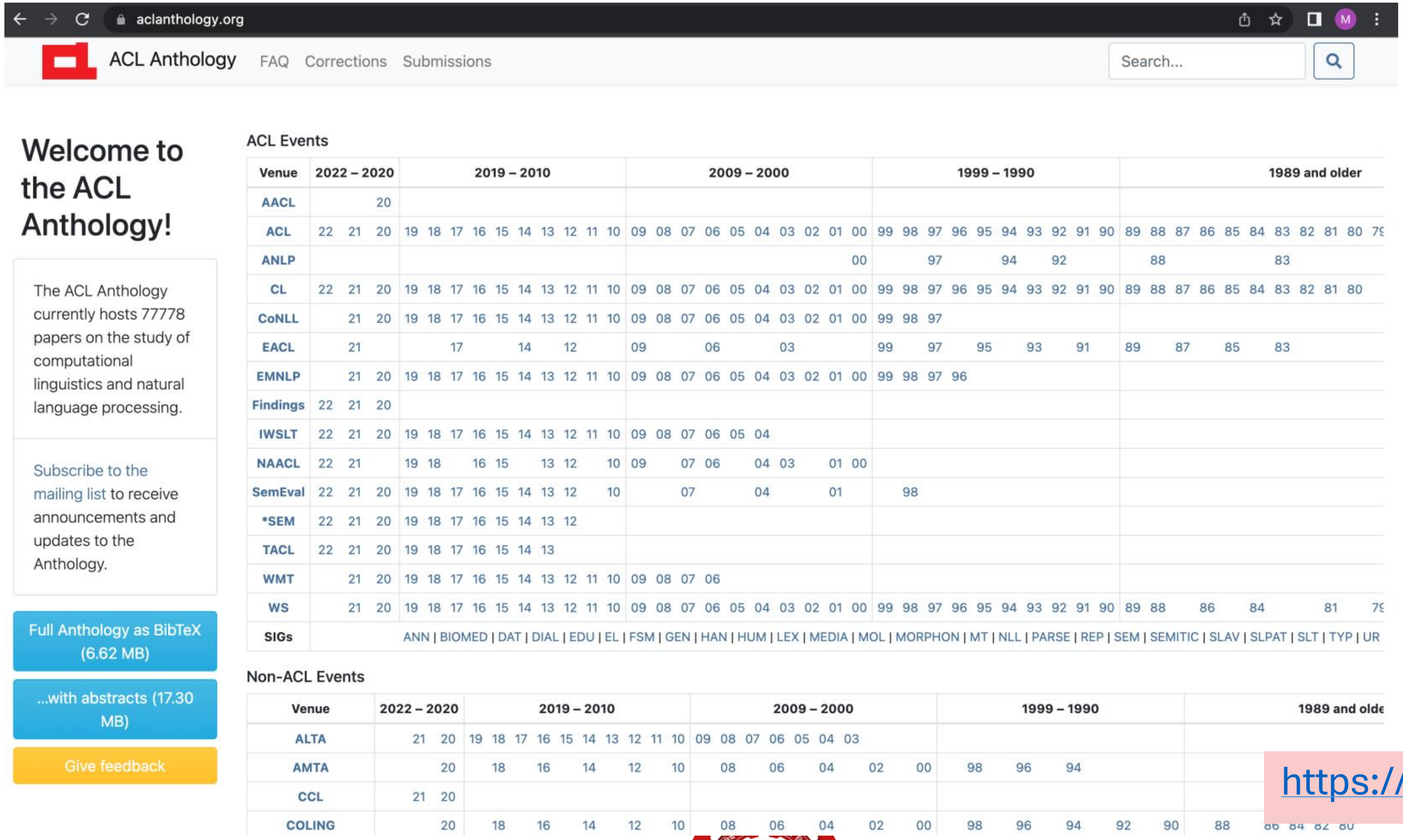
https://www.amazon.in/dp/936386474X/



Other Reference Materials

- Reference Books (optional reading)
  - Speech and Language Processing, Dan Jurafsky and James H. Martin https://web.stanford.edu/~jurafsky/slp3/
  - Foundations of Statistical Natural Language Processing, Chris Manning and Hinrich Schütze
  - Natural Language Processing, Jacob Eisenstein https://github.com/jacobeisenstein/gt-nlp-class/blob/master/notes/eisenstein-nlp-notes.pdf
  - A Primer on Neural Network Models for Natural Language Processing, Yoav Goldberg http://u.cs.biu.ac.il/~yogo/nnlp.pdf
- Journals
  - Computational Linguistics, TACL, JMLR, TMLR, etc.
- Conferences
  - ACL, EMNLP, NAACL, AAAI, ICML, NeurIPS, ICLR, WWW, KDD, SIGIR, etc.

Research Papers Repository



aclanthology.org

ACL Anthology FAQ Corrections Submissions Search...

Welcome to the ACL Anthology!

The ACL Anthology currently hosts 77778 papers on the study of computational linguistics and natural language processing.

Subscribe to the mailing list to receive announcements and updates to the Anthology.

Full Anthology as BibTeX (6.62 MB)

...with abstracts (17.30 MB)

Give feedback

ACL Events

Venue	2022 – 2020	2019 – 2010	2009 – 2000	1999 – 1990	1989 and older
AACL	20				
ACL	22 21 20	19 18 17 16 15 14 13 12 11 10	09 08 07 06 05 04 03 02 01 00	99 98 97 96 95 94 93 92 91 90	89 88 87 86 85 84 83 82 81 80 79
ANLP			00	97 94 92	88 83
CL	22 21 20	19 18 17 16 15 14 13 12 11 10	09 08 07 06 05 04 03 02 01 00	99 98 97 96 95 94 93 92 91 90	89 88 87 86 85 84 83 82 81 80
CoNLL	21 20	19 18 17 16 15 14 13 12 11 10	09 08 07 06 05 04 03 02 01 00	99 98 97	
EACL	21	17 14 12	09 06 03	99 97 95 93 91	89 87 85 83
EMNLP	21 20	19 18 17 16 15 14 13 12 11 10	09 08 07 06 05 04 03 02 01 00	99 98 97 96	
Findings	22 21 20				
IWSLT	22 21 20	19 18 17 16 15 14 13 12 11 10	09 08 07 06 05 04		
NAACL	22 21	19 18 16 15 13 12 10	09 07 06 04 03 01 00		
SemEval	22 21 20	19 18 17 16 15 14 13 12 10	07 04 01	98	
*SEM	22 21 20	19 18 17 16 15 14 13 12			
TACL	22 21 20	19 18 17 16 15 14 13			
WMT	21 20	19 18 17 16 15 14 13 12 11 10	09 08 07 06		
WS	21 20	19 18 17 16 15 14 13 12 11 10	09 08 07 06 05 04 03 02 01 00	99 98 97 96 95 94 93 92 91 90	89 88 86 84 81 79
SIGs	ANN | BIOMED | DAT | DIAL | EDU | EL | FSM | GEN | HAN | HUM | LEX | MEDIA | MOL | MORPHON | MT | NLL | PARSE | REP | SEM | SEMITIC | SLAV | SLPAT | SLT | TYP | UR				

Non-ACL Events

Venue	2022 – 2020	2019 – 2010	2009 – 2000	1999 – 1990	1989 and olde
ALTA	21 20	19 18 17 16 15 14 13 12 11 10	09 08 07 06 05 04 03		
AMTA	20	18 16 14 12 10	08 06 04 02 00	98 96 94	
CCL	21 20				
COLING	20	18 16 14 12 10	08 06 04 02 00	98 96 94 92 90	88 86 84 82 80

https://aclanthology.org/

Research Papers Repository

arXiv.org > cs > cs.CL

Computation and Language

Authors and titles for recent submissions

- Wed, 19 Aug 2020
- Tue, 18 Aug 2020
- Mon, 17 Aug 2020
- Fri, 14 Aug 2020
- Thu, 13 Aug 2020

[total of 84 entries: **1-25** | 26-50 | 51-75 | 76-84]
[showing 25 entries per page: fewer | more | all]

Wed, 19 Aug 2020

[1] arXiv:2008.07905 [pdf, other]
Glancing Transformer for Non-Autoregressive Neural Machine Translation
Lihua Qian, Hao Zhou, Yu Bao, Mingxuan Wang, Lin Qiu, Weinan Zhang, Yong Yu, Lei Li
Comments: 11 pages, 3 figures, 4 tables
Subjects: **Computation and Language (cs.CL)**

[2] arXiv:2008.07880 [pdf, other]
COVID-SEE: Scientific Evidence Explorer for COVID-19 Related Research
Karin Verspoor, Simon Šuster, Yulia Otmakhova, Shevon Mendis, Zenan Zhai, Biaoyan Fang, Jey Han Lau, Timothy Bal
Comments: COVID-SEE is available at this http URL
Subjects: **Computation and Language (cs.CL)**; Information Retrieval (cs.IR)

[3] arXiv:2008.07772 [pdf, other]
Very Deep Transformers for Neural Machine Translation
Xiaodong Liu, Kevin Duh, Liyuan Liu, Jianfeng Gao
Comments: 6 pages, 3 figures and 3 tables
Subjects: **Computation and Language (cs.CL)**

[4] arXiv:2008.07723 [pdf, other]
NASE: Learning Knowledge Graph Embedding for Link Prediction via Neural Architecture Search
Xiaoyu Kou, Bingfeng Luo, Huang Hu, Yan Zhang
Comments: Accepted by CIKM 2020, short paper
Subjects: **Computation and Language (cs.CL)**

https://arxiv.org/list/cs.CL/recent

Acknowledgements (Non-exhaustive List)

- Advanced NLP, Graham Neubig http://www.phontron.com/class/anlp2022/
- Advanced NLP, Mohit Iyyer https://people.cs.umass.edu/~miyyer/cs685/
- NLP with Deep Learning, Chris Manning, http://web.stanford.edu/class/cs224n/
- Understanding Large Language Models, Danqi Chen https://www.cs.princeton.edu/courses/archive/fall22/cos597G/
- Natural Language Processing, Greg Durrett https://www.cs.utexas.edu/~gdurrett/courses/online-course/materials.html
- Large Language Models: https://stanford-cs324.github.io/winter2022/
- Natural Language Processing at UMBC, https://laramartin.net/NLP-class/
- Computational Ethics in NLP, https://demo.clab.cs.cmu.edu/ethical_nlp/
- Self-supervised models, CS 601.471/671: Self-supervised Models (jhu.edu)
- WING.NUS Large Language Models, https://wing-nus.github.io/cs6101/
- And many more…

What is a Language Model (LM)?

Language Model gives the probability distribution over a sequence of tokens.

What is a Language Model (LM)?

Language Model gives the probability distribution over a sequence of tokens.

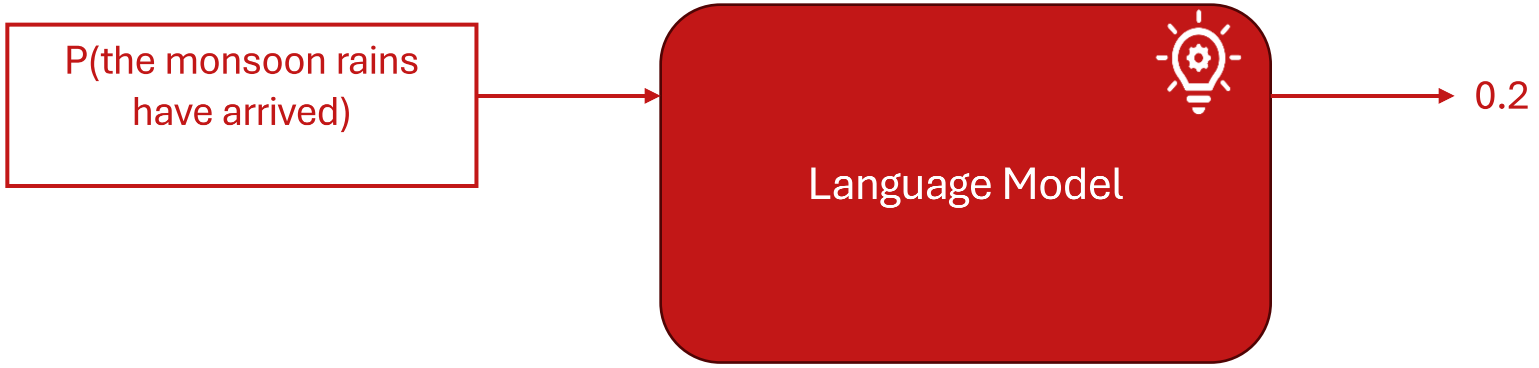
Language Model

Vocabulary

V = {arrived, delhi, have, is, monsoon, rains, the}

What is a Language Model (LM)?

Language Model gives the probability distribution over a sequence of tokens.

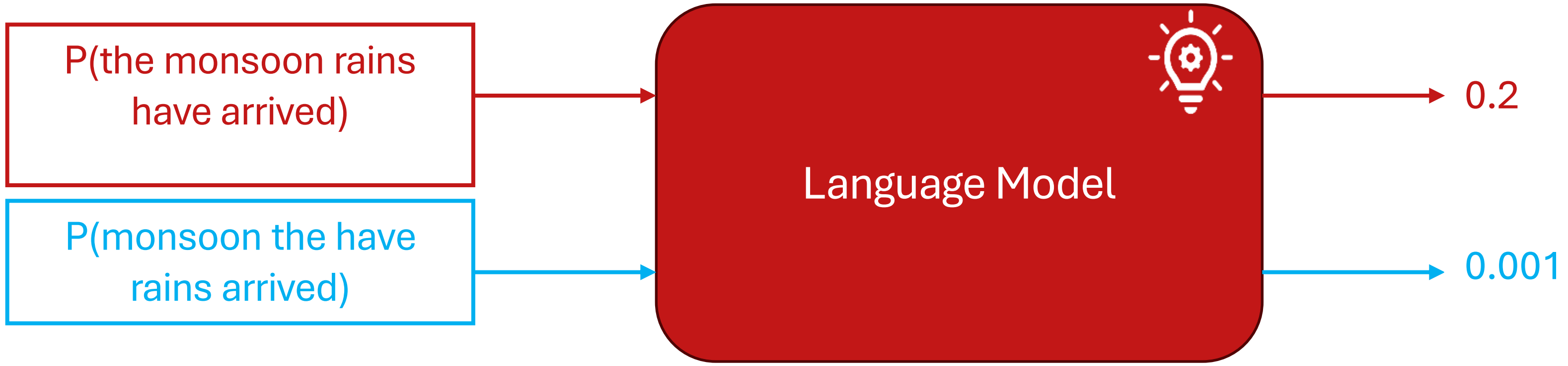


Vocabulary

V = {arrived, delhi, have, is, monsoon, rains, the}

What is a Language Model (LM)?

Language Model gives the probability distribution over a sequence of tokens.



Vocabulary

V = {arrived, delhi, have, is, monsoon, rains, the}

LMs can 'Generate' Text !

- Consider a sequence of tokens $\{x_1, x_2, \dots, x_L\}$,where $x_1, x_2, \dots, x_L$ are in vocabulary V
- Notation: $P(x_1, x_2, \dots, x_L) = P(x_{1:L})$
- Using the **chain rule of probability:**

$$P(x_{1:L}) = P(x_1).P(x_2|x_1).P(x_3|x_1, x_2)\dots P(x_L|x_{L-1}) = \prod_{i=1}^{L} P(x_i|x_{1:i-1})$$

LMs can 'Generate' Text !

- Consider a sequence of tokens $\{x_1, x_2, \ldots, x_L\}$,where $x_1, x_2, \ldots, x_L$ are in vocabulary V
- Notation: $P(x_1, x_2, \ldots, x_L) = P(x_{1:L})$
- Using the **chain rule of probability:**

$$P(x_{1:L}) = P(x_1).P(x_2|x_1).P(x_3|x_1,x_2)\ldots P(x_L|x_{L-1}) = \prod_{i=1}^{L} P(x_i|x_{1:i-1})$$

Vocabulary

V = {arrived, delhi, have, is, monsoon, rains, the}

Given input 'the monsoon rains have' , LM can calculate

$P(x_i \mid \text{the monsoon rains have})$, $\forall\, x_i \in V$

LMs can 'Generate' Text !

- Consider a sequence of tokens $\{x_1, x_2, \dots, x_L\}$,where $x_1, x_2, \dots, x_L$ are in vocabulary V
- Notation: $P(x_1, x_2, \dots, x_L) = P(x_{1:L})$
- Using the **chain rule of probability:**

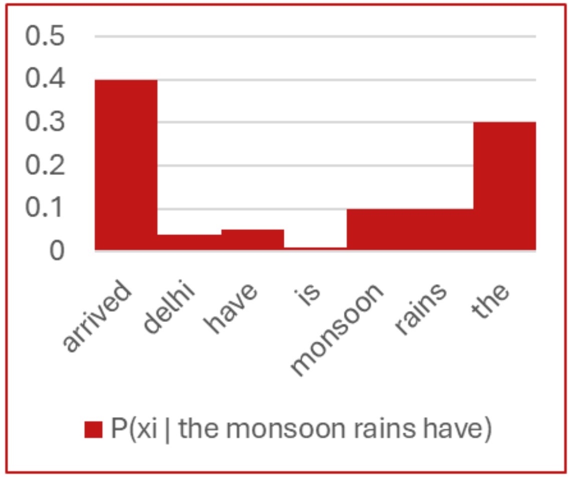
$$P(x_{1:L}) = P(x_1).P(x_2|x_1).P(x_3|x_1,x_2) \dots P(x_L|x_{L-1}) = \prod_{i=1}^{L} P(x_i|x_{1:i-1})$$

Vocabulary

V = {arrived, delhi, have, is, monsoon, rains, the}

Given input 'the monsoon rains have' , LM can calculate

$P(x_i \mid \text{the monsoon rains have})$, $\forall\, x_i \in V$



LMs can 'Generate' Text !

- Consider a sequence of tokens $\{x_1, x_2, \ldots, x_L\}$,where $x_1, x_2, \ldots, x_L$ are in vocabulary V
- Notation: $P(x_1, x_2, \ldots, x_L) = P(x_{1:L})$
- Using the **chain rule of probability:**

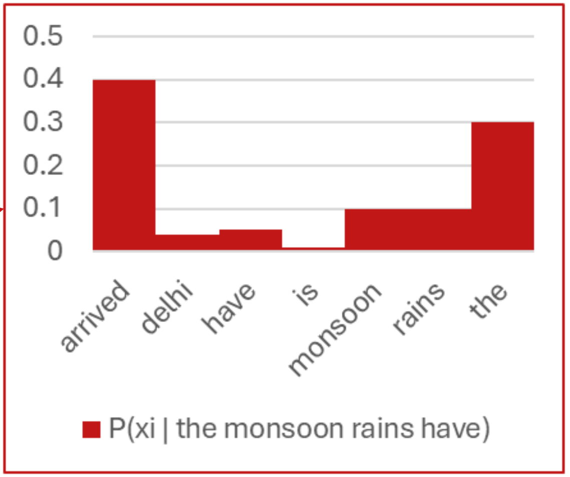
$$P(x_{1:L}) = P(x_1).P(x_2|x_1).P(x_3|x_1,x_2)\ldots P(x_L|x_{L-1}) = \prod_{i=1}^{L} P(x_i|x_{1:i-1})$$

Vocabulary

V = {arrived, delhi, have, is, monsoon, rains, the}

Given input 'the monsoon rains have' , LM can calculate $P(x_i \mid \text{the monsoon rains have})$, $\forall\, x_i \in V$

For generation, next token is sampled from this probability distribution



LMs can 'Generate' Text !

- Consider a sequence of tokens $\{x_1, x_2, \dots, x_L\}$,where $x_1, x_2, \dots, x_L$ are in vocabulary V
- Notation: $P(x_1, x_2, \dots, x_L) = P(x_{1:L})$
- Using the **chain rule of probability:**

$$P(x_{1:L}) = P(x_1).P(x_2|x_1).P(x_3|x_1, x_2) \dots P(x_L|x_{L-1}) = \prod_{i=1}^{L} P(x_i|x_{1:i-1})$$

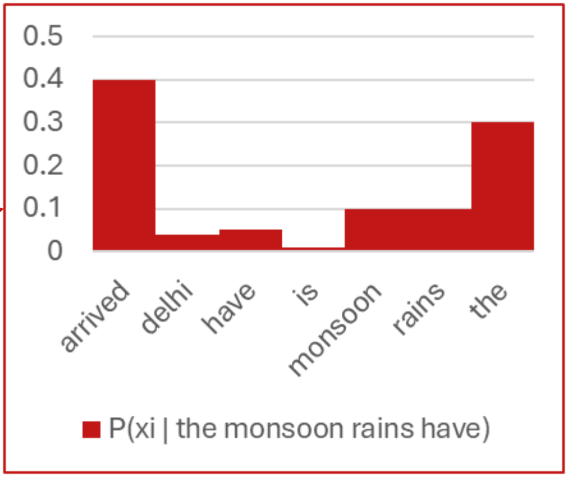
Vocabulary

V = {arrived, delhi, have, is, monsoon, rains, the}

Given input 'the monsoon rains have' , LM can calculate
P(x_i | the monsoon rains have) , $\forall x_i \in V$

For generation, next token is sampled from this probability distribution

$$\boxed{x_i \sim P(x_i \mid x_{1:i-1})}$$



LMs can ‘Generate’ Text !

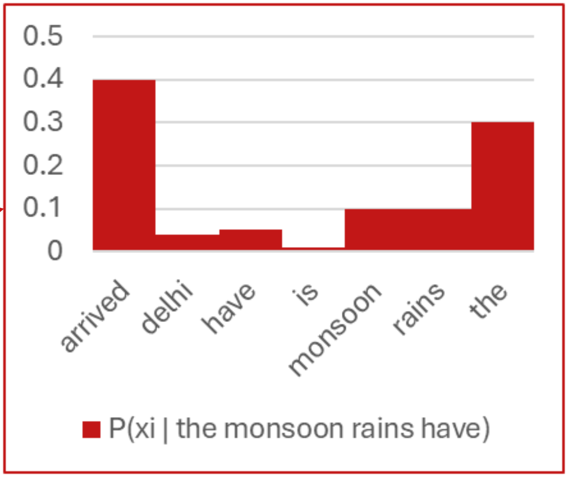
- Consider a sequence of tokens $\{x_1, x_2, \dots, x_L\}$,where $x_1, x_2, \dots, x_L$ are in vocabulary V
- Notation: $P(x_1, x_2, \dots, x_L) = P(x_{1:L})$
- Using the **chain rule of probability:**

$$P(x_{1:L}) = P(x_1).P(x_2|x_1).P(x_3|x_1,x_2)\dots P(x_L|x_{L-1}) = \prod_{i=1}^{L} P(x_i|x_{1:i-1})$$

Vocabulary

V = {arrived, delhi, have, is, monsoon, rains, the}

Given input ‘the monsoon rains have’ , LM can calculate $P(x_i \mid \text{the monsoon rains have})$, $\forall\, x_i \in V$



For generation, next token is sampled from this probability distribution

$x_i \sim P(x_i \mid x_{1:i-1})$

Auto-regressive LMs calculate this distribution efficiently, e.g. using ‘Deep’ Neural Networks

'Large' Language Models

The 'Large' in terms of model's size (# parameters) and massive size of training dataset.

Model	Organization	Date	Size (# params)
ELMo	AI2	Feb 2018	94,000,000
GPT	OpenAI	Jun 2018	110,000,000
BERT	Google	Oct 2018	340,000,000
XLM	Facebook	Jan 2019	655,000,000
GPT-2	OpenAI	Mar 2019	1,500,000,000
RoBERTa	Facebook	Jul 2019	355,000,000
Megatron-LM	NVIDIA	Sep 2019	8,300,000,000
T5	Google	Oct 2019	11,000,000,000
Turing-NLG	Microsoft	Feb 2020	17,000,000,000
GPT-3	OpenAI	May 2020	175,000,000,000
Megatron-Turing NLG	Microsoft, NVIDIA	Oct 2021	530,000,000,000
Gopher	DeepMind	Dec 2021	280,000,000,000

Model sizes have increased by an order of **5000x** over just the last 4 years !!!

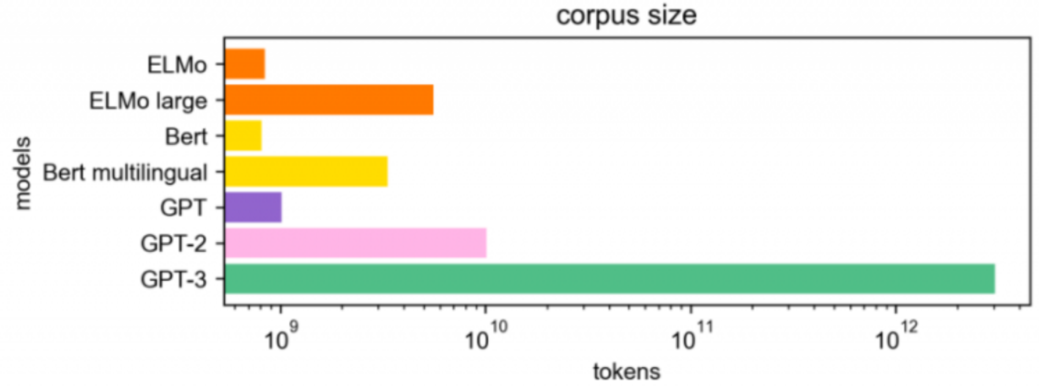


Image source: https://hellofuture.orange.com/en/the-gpt-3-language-model-revolution-or-evolution/

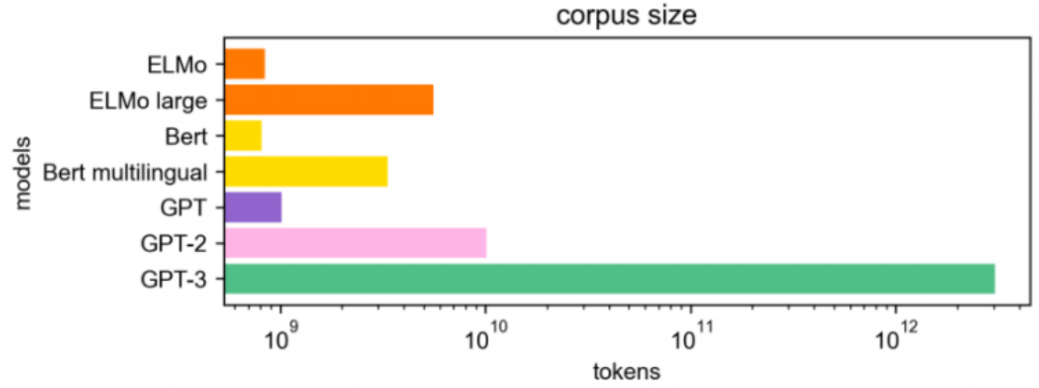
‘Large’ Language Models

The ‘Large’ in terms of model's size (# parameters) and massive size of training dataset.

Model	Organization	Date	Size (# params)
ELMo	AI2	Feb 2018	94,000,000
GPT	OpenAI	Jun 2018	110,000,000
BERT	Google	Oct 2018	340,000,000
XLM	Facebook	Jan 2019	655,000,000
GPT-2	OpenAI	Mar 2019	1,500,000,000
RoBERTa	Facebook	Jul 2019	355,000,000
Megatron-LM	NVIDIA	Sep 2019	8,300,000,000
T5	Google	Oct 2019	11,000,000,000
Turing-NLG	Microsoft	Feb 2020	17,000,000,000
GPT-3	OpenAI	May 2020	175,000,000,000
Megatron-Turing NLG	Microsoft, NVIDIA	Oct 2021	530,000,000,000
Gopher	DeepMind	Dec 2021	280,000,000,000

Image source: https://hellofuture.orange.com/en/the-gpt-3-language-model-revolution-or-evolution/

Model sizes have increased by an order of **5000x** over just the last 4 years !!!



Other recent models: PaLM (540B), OPT (175B), BLOOM (176B), Gemini-Ultra (1.56T), GPT-4 (1.76T)

Disclaimer: For API-based models like GPT-4/Gemini-Ultra, the number of parameters are not announced officially – these are rumored numbers as on the web

LLMs in AI Landscape

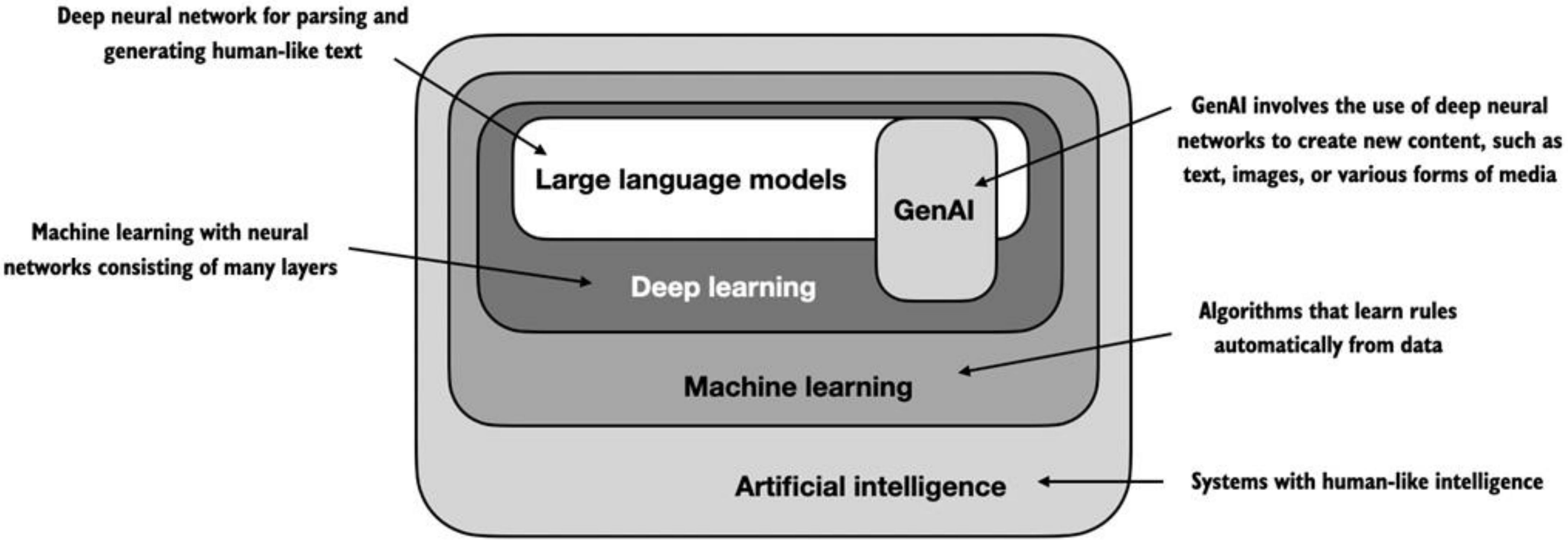


Image source: https://www.manning.com/books/build-a-large-language-model-from-scratch

Evolution of (L)LMs

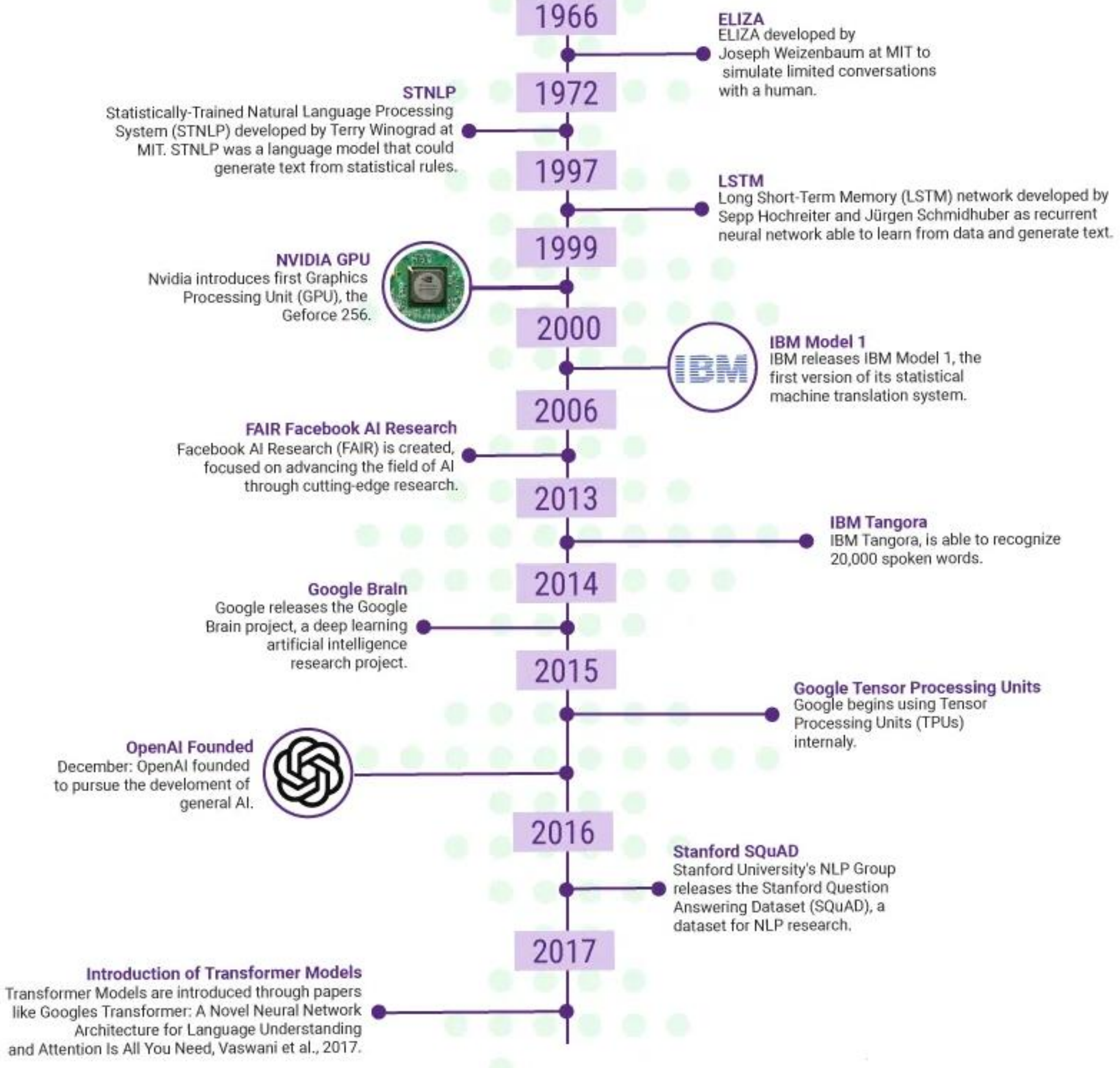


Image source: https://synthedia.substack.com/p/a-timeline-of-large-language-model

Post-Transformers Era

The LLM Race

Google Designed Transformers: But Could it Take Advantage?

Transformers (2017)

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Google Designed Transformers: But Could it Take Advantage?

Transformers (2017)

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser* Google Brain lukaszkaiser@google.com
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

BERT (2018)

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin **Ming-Wei Chang** **Kenton Lee** **Kristina Toutanova**
Google AI Language
{jacobdevlin,mingweichang,kentonl,kristout}@google.com

Google Designed Transformers: But Could it Take Advantage?

Transformers (2017)

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser* Google Brain lukaszkaiser@google.com
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

BERT (2018)

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova
Google AI Language
{jacobdevlin,mingweichang,kentonl,kristout}@google.com

The beginning of use of Transformer as Language Representation Models.

BERT achieved SOTA on 11 NLP tasks.

Google Designed Transformers: But Could it Take Advantage?

Transformers (2017)

BERT (2018)

DistilBERT, TinyBERT, MobileBERT

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser* Google Brain lukaszkaiser@google.com
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin **Ming-Wei Chang** **Kenton Lee** **Kristina Toutanova**
Google AI Language
{jacobdevlin,mingweichang,kentonl,kristout}@google.com

The beginning of use of Transformer as Language Representation Models.

BERT achieved SOTA on 11 NLP tasks.

However, someone was waiting for the right opportunity!!

Guess Who?

However, someone was waiting for the right opportunity!!

OpenAI

OpenAI Started Pushing the Frontier

OpenAI

Improving Language Understanding by Generative Pre-Training

Alec Radford
OpenAI
alec@openai.com

Karthik Narasimhan
OpenAI
karthikn@openai.com

Tim Salimans
OpenAI
tim@openai.com

Ilya Sutskever
OpenAI
ilyasu@openai.com

OpenAI Started Pushing the Frontier

GPT (2018)

OpenAI

Improving Language Understanding by Generative Pre-Training

Alec Radford	Karthik Narasimhan	Tim Salimans	Ilya Sutskever
OpenAI	OpenAI	OpenAI	OpenAI
alec@openai.com	karthikn@openai.com	tim@openai.com	ilyasu@openai.com

OpenAI Started Pushing the Frontier

GPT (2018)

OpenAI

Improving Language Understanding by Generative Pre-Training

Alec Radford OpenAI alec@openai.com

Karthik Narasimhan OpenAI karthikn@openai.com

Tim Salimans OpenAI tim@openai.com

Ilya Sutskever OpenAI ilyasu@openai.com

- Use of decoder-only architecture
- The idea of generative pre-training over large corpus

The Beginning of Scale

GPT-2 (2019)

Language Models are Unsupervised Multitask Learners

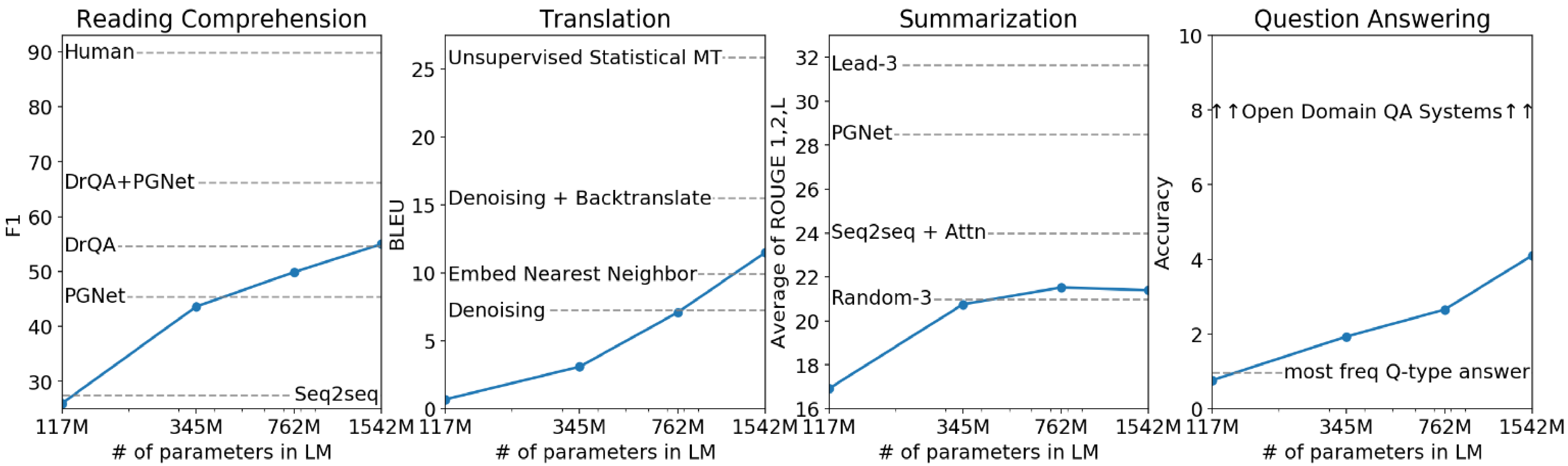
Alec Radford [*1] **Jeffrey Wu** [*1] **Rewon Child** [1] **David Luan** [1] **Dario Amodei** [**1] **Ilya Sutskever** [**1]

- GPT-1 (117 M) → GPT-2 (1.5 B) **13x increase in # parameters**
- Minimal changes (some LayerNorms added, modified weight initialization)
- Increase in context length: GPT-1 (512 tokens) → GPT-2 (1024 tokens)

The Beginning of Scale

GPT-2 (2019)

Performance boosts across tasks



What Was Google Developing Parallelly?

T5 (2019)

Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer

Colin Raffel*	CRAFFEL@GMAIL.COM
Noam Shazeer*	NOAM@GOOGLE.COM
Adam Roberts*	ADAROB@GOOGLE.COM
Katherine Lee*	KATHERINELEE@GOOGLE.COM
Sharan Narang	SHARANNARANG@GOOGLE.COM
Michael Matena	MMATENA@GOOGLE.COM
Yanqi Zhou	YANQIZ@GOOGLE.COM
Wei Li	MWEILI@GOOGLE.COM
Peter J. Liu	PETERJLIU@GOOGLE.COM

Google, Mountain View, CA 94043, USA

What Was Google Developing Parallelly?

T5 (2019)

Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer

Colin Raffel* CRAFFEL@GMAIL.COM
Noam Shazeer* NOAM@GOOGLE.COM
.COM
.COM
.COM
.COM
.COM
.COM
.COM

- Similar broader goal of converting all text-based language problems into a text-to-text format.
- Used **Encoder-Decoder Architecture**.
- Pre-training strategy differs from GPT
  - Strategy more similar to BERT

Google, Mountain View, CA 94043, USA

Was It Only Google vs OpenAI? Where did **Meta** Stand?

Was It Only Google vs OpenAI?
Where did **Meta** Stand?

RoBERTa
(2019)

RoBERTa: A Robustly Optimized BERT Pretraining Approach

Yinhan Liu[*§] **Myle Ott**[*§] **Naman Goyal**[*§] **Jingfei Du**[*§] **Mandar Joshi**[†]
Danqi Chen[§] **Omer Levy**[§] **Mike Lewis**[§] **Luke Zettlemoyer**[†§] **Veselin Stoyanov**[§]

[†] Paul G. Allen School of Computer Science & Engineering,
University of Washington, Seattle, WA
{mandar90,lsz}@cs.washington.edu

[§] Facebook AI
{yinhanliu,myleott,naman,jingfeidu,
danqi,omerlevy,mikelewis,lsz,ves}@fb.com

Was It Only Google vs OpenAI? Where did **Meta** Stand?

RoBERTa (2019)

RoBERTa: A Robustly Optimized BERT Pretraining Approach

Yinhan Li **Iandar Joshi**[†]
Danqi Chen[§] **Veselin Stoyanov**[§]
ng,
fb.com

- Replication study of BERT pretraining
- Measured the impact of many key hyperparameters and training data size.
- **Found that BERT was significantly undertrained**, and can match or exceed the performance of every model published after it.

Was It Only Google vs OpenAI? Where did **Meta** Stand?

RoBERTa (2019)

RoBERTa: A Robustly Optimized BERT Pretraining Approach

Yinhan Li ... Mandar Joshi[†]
Danqi Chen[§] ... Veselin Stoyanov[§]
...ng,
...fb.com

- Replication study of BERT pretraining
- Measured the impact of many key hyperparameters and training data size.
- **Found that BERT was significantly undertrained**, and can match or exceed the performance of every model published after it.

XLM (2019)

Cross-lingual Language Model Pretraining

Guillaume Lample* Facebook AI Research Sorbonne Universités glample@fb.com

Alexis Conneau* Facebook AI Research Université Le Mans aconneau@fb.com

Was It Only Google vs OpenAI? Where did **Meta** Stand?

RoBERTa (2019)

RoBERTa: A Robustly Optimized BERT Pretraining Approach

Yinhan Li ... Iandar Joshi[†]
Danqi Chen[§] ... Veselin Stoyanov[§]
ng,
fb.com

- Replication study of BERT pretraining
- Measured the impact of many key hyperparameters and training data size.
- **Found that BERT was significantly undertrained**, and can match or exceed the performance of every model published after it.

XLM (2019)

Cross-lingual Language Model Pretraining

Guilla ... au*
Faceboo ... earch
Sorbon ... Ians
glamp ... .com

- Proposed methods to learn **cross-lingual language models (XLMs)**
- Obtained SOTA on:
  - cross-lingual classification
  - unsupervised and supervised machine translation

OpenAI Continues to Scale

GPT-3 (2020)

Language Models are Few-Shot Learners

Tom B. Brown* Benjamin Mann* Nick Ryder* Melanie Subbiah*

Jared Kaplan† Prafulla Dhariwal Arvind Neelakantan Pranav Shyam Girish Sastry

Amanda Askell Sandhini Agarwal Ariel Herbert-Voss Gretchen Krueger Tom Henighan

Rewon Child Aditya Ramesh Daniel M. Ziegler Jeffrey Wu Clemens Winter

Christopher Hesse Mark Chen Eric Sigler Mateusz Litwin Scott Gray

Benjamin Chess Jack Clark Christopher Berner

Sam McCandlish Alec Radford Ilya Sutskever Dario Amodei

OpenAI

OpenAI Continues to Scale

GPT-3 (2020)

Language Models are Few-Shot Learners

Tom B. Brown* Benjamin Mann* Nick Ryder* Melanie Subbiah*

Jared Kaplan† Prafu ranav Shyam Girish Sastry

Amanda Askell Sandl hen Krueger Tom Henighan

Rewon Child Adi rey Wu Clemens Winter

Christopher Hesse Mark Chen Eric Sigler Mateusz Litwin Scott Gray

Benjamin Chess Jack Clark Christopher Berner

Sam McCandlish Alec Radford Ilya Sutskever Dario Amodei

OpenAI

175 B parameters !

OpenAI Continues to Scale

GPT-3 (2020)

Language Models are Few-Shot Learners

Tom B. Brown* Benjamin Mann* Nick Ryder* Melanie Subbiah*

Jared Kaplan† Prafu ranav Shyam Girish Sastry

Amanda Askell Sandh hen Krueger Tom Henighan

Rewon Child Adit rey Wu Clemens Winter

Christopher H win Scott Gray

Benja pher Berner

Sam McCa Dario Amodei

OpenAI

175 B parameters !

OpenAI stops open-sourcing!!

Google Starts Scaling too (But is it Late) !

PaLM (2022)

PaLM: Scaling Language Modeling with Pathways

Aakanksha Chowdhery* Sharan Narang* Jacob Devlin*
Maarten Bosma Gaurav Mishra Adam Roberts Paul Barham
Hyung Won Chung Ch ... rker Schuh Kensen Shi
Sasha Tsvyashchenk ... ker Barnes Yi Tay
Noam Shazeer‡ Vino ... Du Ben Hutchinson
Reiner Pope Jan ... ard Guy Gur-Ari
Pengcheng Yin To ... mawat Sunipa Dev
Henryk Michalewski Xavier Garcia Vedant Misra Kevin Robinson Liam Fedus
Denny Zhou Daphne Ippolito David Luan‡ Hyeontaek Lim Barret Zoph
Alexander Spiridonov Ryan Sepassi David Dohan Shivani Agrawal Mark Omernick
Andrew M. Dai Thanumalayan Sankaranarayana Pillai Marie Pellat Aitor Lewkowycz
Erica Moreira Rewon Child Oleksandr Polozov† Katherine Lee Zongwei Zhou
Xuezhi Wang Brennan Saeta Mark Diaz Orhan Firat Michele Catasta† Jason Wei
Kathy Meier-Hellstern Douglas Eck Jeff Dean Slav Petrov Noah Fiedel

Google Research

540 B parameters !

Google Starts Scaling too (But is it Late) !

PaLM (2022)

PaLM: Scaling Language Modeling with Pathways

Aakanksha Chowdhery* Sharan Narang* Jacob Devlin*
Maarten Bosma Gaurav Mishra Adam Roberts Paul Barham
Hyung Won Chung Ch rker Schuh Kensen Shi
Sasha Tsvyashchenk ker Barnes Yi Tay
Noam Shazeer‡ Vino Du Ben Hutchinson
Reiner Pope Jan ard Guy Gur-Ari
Pengcheng Yin To mawat Sunipa Dev
Henryk Michalewski Xavier Garcia Vedant Misra Kevin Robinson Liam Fedus
Denny Zho Barret Zoph
Alexander Spirid Mark Omernick
Andrew M. Dai Aitor Lewkowycz
Erica Moreira Zongwei Zhou
Xuezhi Wang E asta† Jason Wei
Kathy Meie Noah Fiedel
Google Research

540 B parameters !

Google follows OpenAI in stopping open-sourcing !

It's now the "LLM Race"

2021-2022: A Flurry of LLMs

Microsoft
NVIDIA
Megatron-Turing NLG

Gopher

DeepMind
Chinchilla AI

LaMDA

CODEGEN

OpenAI
Codex

Google AI
MINERVA
MODEL FOR QUANTITATIVE REASONING PROBLEMS SOLVING

Meta Promotes Open-sourcing !

Meta Promotes Open-sourcing !

OPT (2022)

OPT: Open Pre-trained Transformer Language Models

Susan Zhang*, Stephen Roller*, Naman Goyal*,
Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li,
Xi Victoria Lin, Todor Mihaylov, Myle Ott†, Sam Shleifer†, Kurt Shuster, Daniel Simig,
Punit Singh Koura, Anjali Sridhar, Tianlu Wang, Luke Zettlemoyer

Meta AI

{susanz,roller,naman}@fb.com

Meta Promotes Open-sourcing !

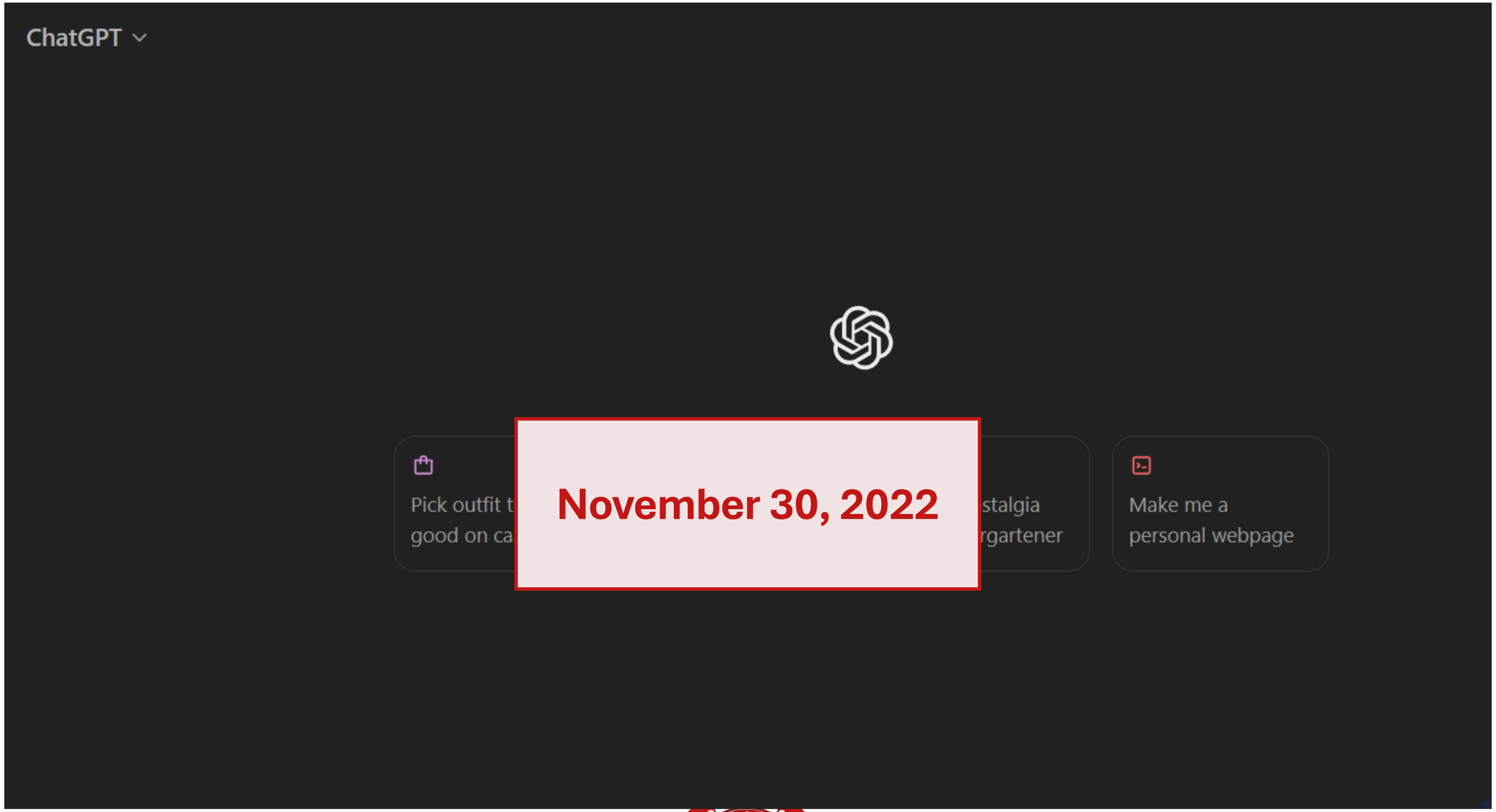
OPT (2022)

OPT: Open Pre-trained Transformer Language Models

Susan Zhang*, Stephen Roller*, Naman Goyal*
Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li,
Xi Victoria Lin, T… t Shuster, Daniel Simig,
Punit Sin… ke Zettlemoyer

- A suite of decoder-only pre-trained transformers ranging from 125M to 175B parameters
- **Open-sourced !!!**

The ChatGPT Moment



2023: The Year of Rapid Pace

Feb, 2023: **Google** releases **Bard**

2023: The Year of Rapid Pace

Feb, 2023: **Google** releases **Bard**

Feb, 2023: **Meta** releases its **LLaMA** family of **open-source models**

2023: The Year of Rapid Pace

Bard

Feb, 2023: **Google** releases **Bard**

Feb, 2023: **Meta** releases its **LLaMA** family of **open-source models**

Claude

BY ANTHROP\C

March, 2023: **Anthropic**, a start-up founded in 2021 by ex-OpenAI researchers, releases **Claude**

2023: The Year of Rapid Pace

Feb, 2023: **Google** releases **Bard**

Feb, 2023: **Meta** releases its **LLaMA** family of **open-source models**

March, 2023: **Anthropic**, a start-up founded in 2021 by ex-OpenAI researchers, releases **Claude**

March, 2023: **OpenAI** releases **GPT-4**

2023: The Year of Rapid Pace

Feb, 2023: **Google** releases **Bard**

Feb, 2023: **Meta** releases its **LLaMA** family of **open-source models**

March, 2023: **Anthropic**, a start-up founded in 2021 by ex-OpenAI researchers, releases **Claude**

March, 2023: **OpenAI** releases **GPT-4**

Sept, 2023: **Mistral AI** releases **Mistral-7B** model

2023: The Year of Rapid Pace

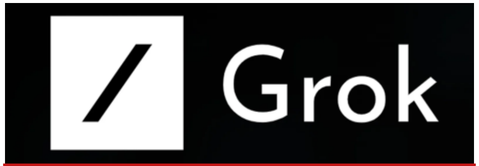
Feb, 2023: **Google** releases **Bard**

Feb, 2023: **Meta** releases its **LLaMA** family of **open-source models**

March, 2023: **Anthropic**, a start-up founded in 2021 by ex-OpenAI researchers, releases **Claude**

March, 2023: **OpenAI** releases **GPT-4**

Sept, 2023: **Mistral AI** releases **Mistral-7B** model



Nov, 2023: **xAI** releases **Grok**

2023: The Year of Rapid Pace

Feb, 2023: **Google** releases **Bard**

Feb, 2023: **Meta** releases its **LLaMA** family of **open-source models**

March, 2023: **Anthropic**, a start-up founded in 2021 by ex-OpenAI researchers, releases **Claude**

March, 2023: **OpenAI** releases **GPT-4**

Sept, 2023: **Mistral AI** releases **Mistral-7B** model

Nov, 2023: **xAI** releases **Grok**

Dec, 2023: **Google** releases **Gemini**

Despite all those advanced models, LLMs were (and probably still are) pretty bad at reasoning tasks!

Since 2024 we are witnessing more advanced reasoning models!

Also called: **“Reasoning Models”**, **“Large Reasoning Models”**, …

- **Interest in reasoning surged after OpenAI released o1**
  - Showed improved performance across reasoning tasks
  - OpenAI didn’t reveal how they improved reasoning
  - **OpenAI claimed**: these models are trained to **“think”** before responding
    - **Classic case of anthropomorphizing LLMs !!!**

- **Open-source LLMs seemed to lag behind in reasoning capabilities after launch of o1 !**

DeepSeek: Open-Source Reasoning LLMs

- **DeepSeek-R1** – first **open-source Large Reasoning Model** to **rival OpenAI o1**
  - Grabbed the AI community's attention because it showed that an **open, low-cost model can match premium, closed systems on tough reasoning tasks while staying efficient enough to run locally.**
- Fast local inference
- **Efficient and cheap training**
  - DeepSeek claimed the final R1 run cost **under USD 6 million on H800 GPUs**, shattering assumptions that GPT-level performance needs nine-figure budgets

All other organizations followed

And here we are in 2025!

RESEARCH

Advanced version of Gemini with Deep Think officially achieves gold-medal standard at the International Mathematical Olympiad

21 JULY 2025

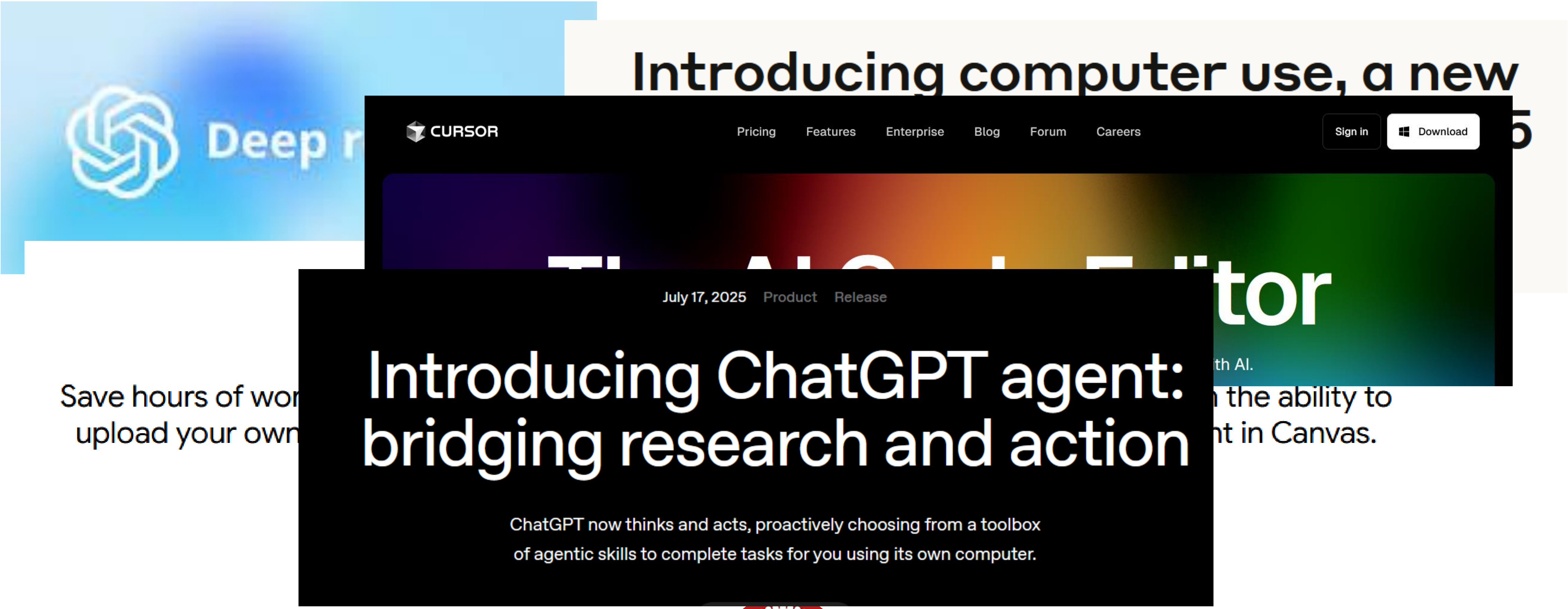
Google and OpenAI's AI models win milestone gold at global math competition

The results marked the first time that AI systems crossed the gold-medal scoring threshold at the International Mathematical Olympiad for high-school students

Published - July 22, 2025 09:55 am IST

REUTERS

The previous year also witnessed an improvement in LLM Agents!



Why Does This Course Exist?

Why Does This Course Exist?

What have changed since 2024? What's new in 2025?

- The fundamentals still remain the same
- However, more **advanced alignment techniques and RL-based post-training strateg**ies are developed for enhancing the reasoning capabilities of models.
- There were also **advancements in the 'Agents' space** – new modelling techniques, better methods and new protocols for handling agents!
- Another axis is **Efficient and Smaller Models**
  - Researchers are interested in developing LLMs which are faster and smaller – without compromising on the quality of outputs and the models' effectiveness.
- These also led to exploring **alternate architectures**, apart from Transformers
  - For example, Diffusion-based LLMs are blazingly fast compared to auto-regressive ones. But their performance are still not up to the mark.
- More and more research are also being done to **demystify the internal working** of LLMs and to understand the mechanisms behind their various phenomena!

We Will Cover These Aspects in 5 Modules

Module-1: Fundamentals

- An overview of the current state in the era of 'LLM Race'
- Details of the **Transformer Architecture**
- Variants of the Transformer Architecture and their **pre-training strategies**
- **Post-training strategies** of modern LLMs (Instruction Tuning, RLHF, etc.)
- **Advanced Alignment Techniques** (PPO, DPO, GRPO, MCTS, PRMs, etc.)

Course Introduction	Introduction to Transformers
Pre-training and Post-training Strategies	Advanced Alignment Techniques

We Will Cover These Aspects in 5 Modules

- Module-2: Efficiency
  - Efficient Design, Training and Inference in Language Models
    - **Mixture of Experts**
    - **Rotary Positional Encoding (RoPE), ALiBi, etc.**
    - **Efficient Attention Mechanisms**
    - **KV Caching**
    - **vLLM**
    - **Efficient Inference Techniques**
  - Various **Parameter Efficient Fine-Tuning (PEFT)** techniques like Prompt Tuning, Prefix Tuning, LoRA, QLoRA, etc,.
  - Various **Model Compression** techniques like model pruning, quantization, etc.

Efficient Design, Training and Inference in LMs

PEFT

Model Compression

We Will Cover These Aspects in 5 Modules

- Module-3: Augmentation & Reasoning
  - **Retrieval-Augmented Language Models**
  - **LLM Agents**
    - Function Calling
    - Design Decisions
    - Protocols (MCP, ACP, A2A, etc.)
  - **Large Reasoning Models (LRMs)**
    - Training to reason via Reinforcement Learning
    - Test-time scaling

Retrieval-Augmented Language Models

LLM Agents

Large Reasoning Models

We Will Cover These Aspects in 5 Modules

- Module-4: Alternate Paradigms
  - **Multimodal Models**
    - Vision Language Models
    - Audio-visual Language Models
  - **Alternative LLM Architectures**
    - State Space Models (SSMs)
    - Diffusion-based LMs
    - Hybrid Models

Multimodal Models

Alternative LLM Architectures

We Will Cover These Aspects in 5 Modules

- Module-5: Miscellaneous
  - **Physics of Language Models**
  - **Interpretability**
    - A peep into the internal workings of LLMs to understand the source of their capabilities
  - A discussion on **ethical issues** and **risks** of LLM usage

Physics of LMs

Interpretability

Ethics and Conclusion

Suggestions (For Effective Learning)

- To understand the concepts clearly, experiment with the models (**Hugging Face** makes life easier).
- Smaller models (like, GPT2) can be run on **Google Colab** / **Kaggle**.
  - Even 7B models can be run with proper quantization.

Always **get your hands dirty !**

LLM Research is all about implementing and experimenting with your ideas.

Suggestions (For Effective Learning)

- To understand the concepts clearly, experiment with the models (**Hugging Face** makes life easier).
- Smaller models (like, GPT2) can be run on **Google Colab** / **Kaggle**.
    - Even 7B models can be run with proper quantization.

Rule of thumb:

Never believe in any hypothesis until your experiments verify it !

REMINDER

You are advised to study the **first 10 lectures (till Lec 6.1)** of the previous year's course playlist before the **next class on August 4**. Otherwise, you will not be able to follow. Here's the link to the playlist:

WE WILL NOT HAVE CLASSES ON JULY 28, 30, 31 DUE TO ACL!

See you all in-person on August 4 !