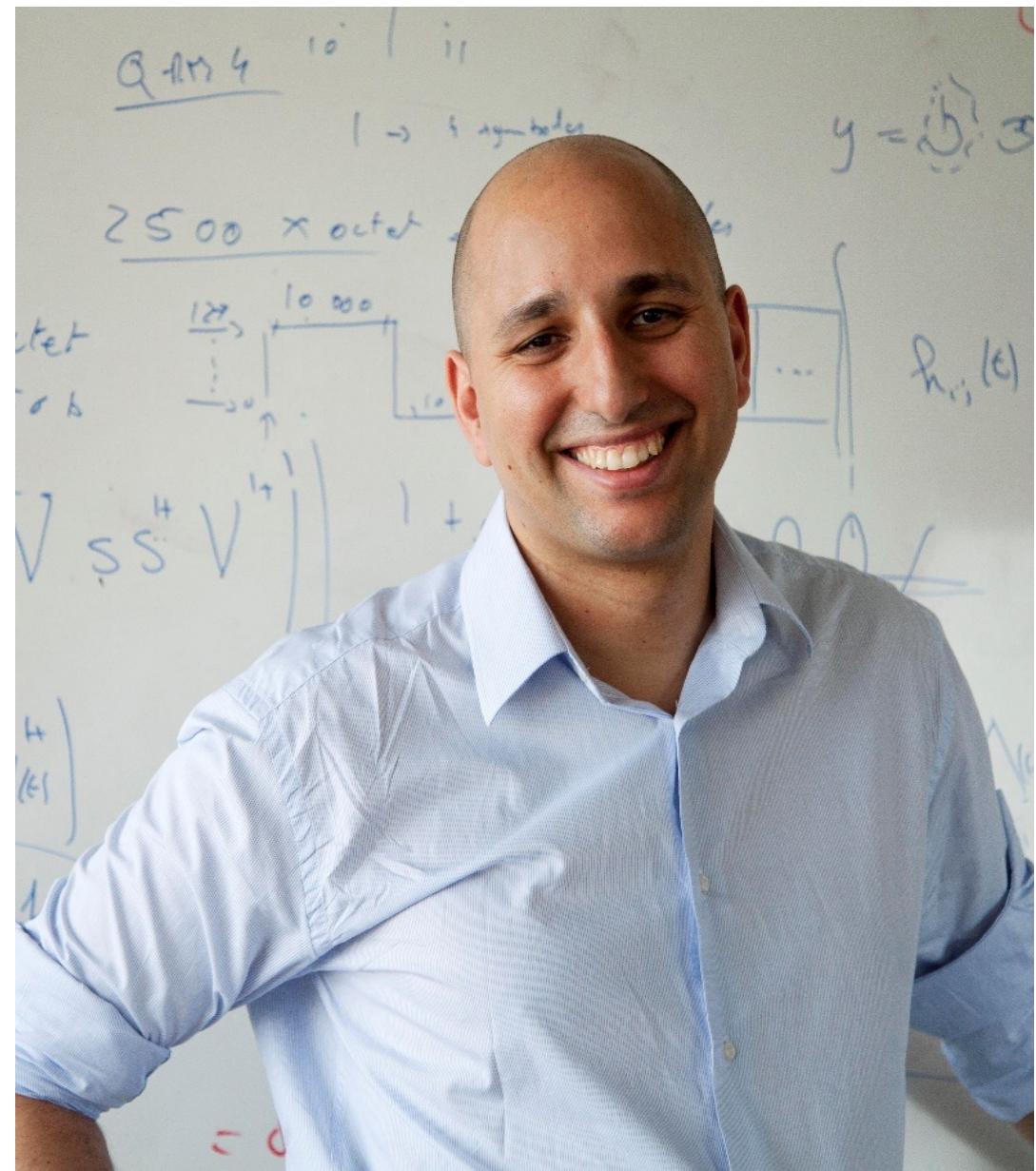


Generative AI: Challenges and Opportunities

Prof. Merouane Debbah

About The Researcher

- Professor & Founding Director, Khalifa University 6G Center
- IEEE, EURASIP and WWRF Fellow
- Citations: 74000+, h-index:119
- More than 40 IEEE Best papers Awards
- More than 50 patents
- IEEE Signal Processing Society Distinguished Industry Speaker (2021-2022)
- Field of Research: AI, Mathematics and Networks



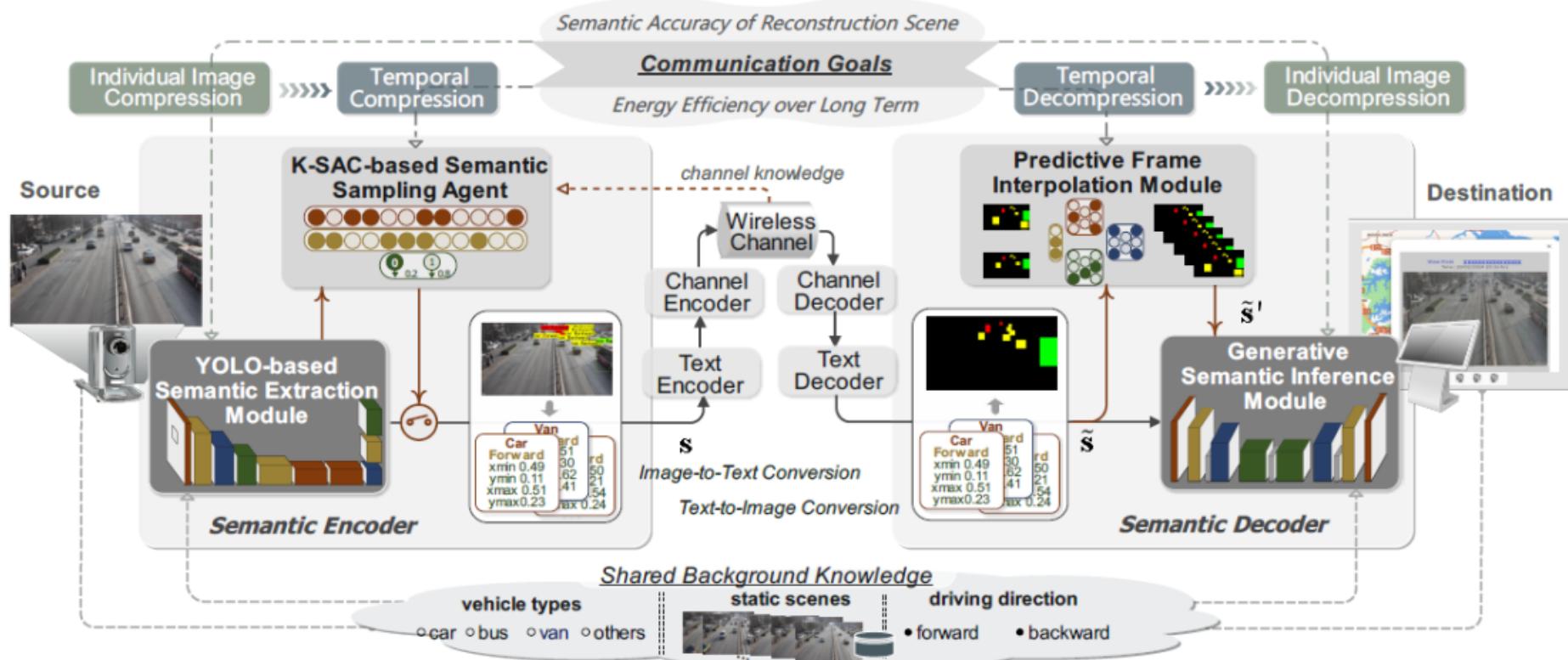
6G: Connected Intelligence

“G” Waves

- 2G: Mobile for Voice
- 3G: Mobile for Data
- 4G: Mobile for Internet
- 5G: Mobile for Things
- 6G : Mobile for Intelligence

	Human	Machine
Display resolution	< 290 ppi	Limits??
Display refresh rate	< 60 fps	
3D effect	Left / right image	
Latency	< 100 ms	
Audible frequency	250 ~ 2000 Hz	
Visible frequency	280 ~ 780 nm	
Viewing angle	Horizontal 200°, vertical 130°	
Senses	5 senses (see, hear, smell, touch, taste)	

Semantic Communication: LLM for Compression



Telecom AI agents: end of apps...

Media | 02-15-2024 | Niels Hafnerichter | 2 Comments

AI phone: Deutsche Telekom wants to free smartphones from apps

 Share  Print  Read out

- Visionary showcase at MWC 2024 shows the world of an app-free AI smartphone
- Digital assistant helps in (almost) all situations in life
- Cooperation with Qualcomm and Brain.ai



Portable Large Language Models – not the iPhone 15 – are the future of the smartphone

Personal AI can redefine the handheld experience and perhaps preserve privacy too

 Mark Pesce

Wed 13 Sep 2023 // 07:38 UTC

COLUMN Smartphone innovation has plateaued. The iPhone 15, launched overnight, has some nice additions. But my iPhone 13 will meet my needs for a while and I won't rush to replace it. My previous iPhone lasted four years.

Before that phone I could justify grabbing Cupertino's annual upgrade. These days, what do we get? The iPhone 15 delivered USB-C, a better camera, and faster wireless charging. It's all nice, but not truly necessary for most users.

Yet smartphones are about to change for the better – thanks to the current wild streak of innovation around AI.

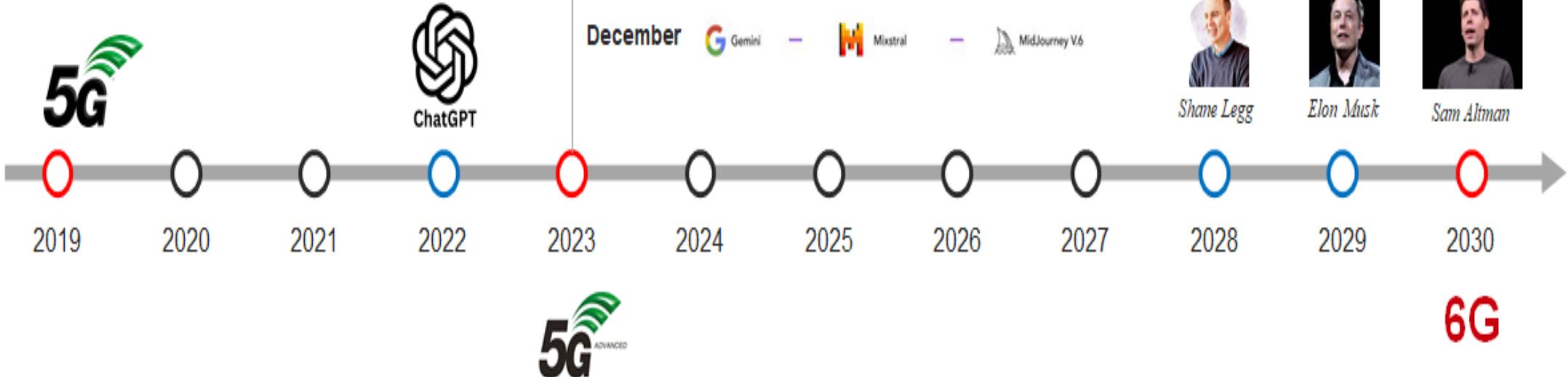
Pretty much everyone with a smartphone can already access the "Big Three" AI chatbots – OpenAI's ChatGPT, Microsoft's Bing Chat and Google's Bard – through an app or browser.

That works well enough. Yet alongside these "general purpose" AI chatbots, a

Around 2030, 6G and AGI will likely arrive at the same time

If an AI system achieves the current level of human performance on all human cognitive tasks that can be proposed, this system could be considered AGI.

— Shane Legg, DeepMind



Talk Outline

- A General Introduction
- Our Work on Generative AI
- Thoughts on next steps in Generative AI

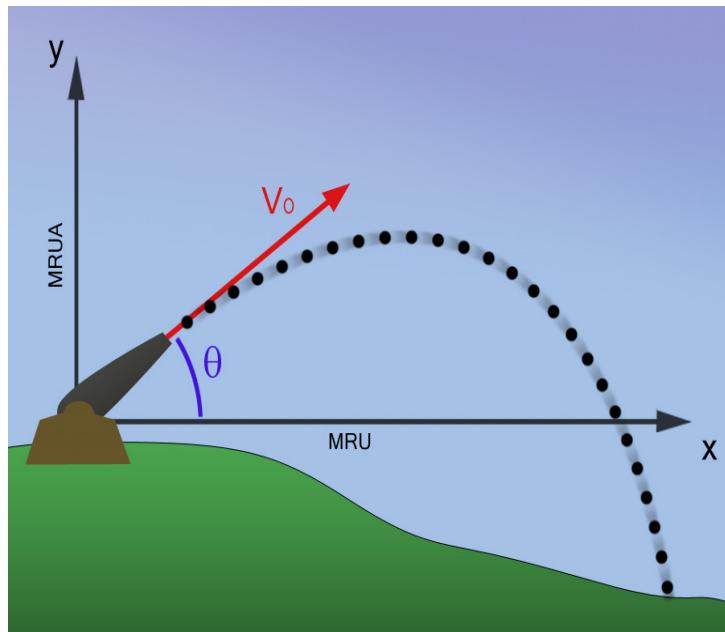
Talk Outline

- A General Introduction

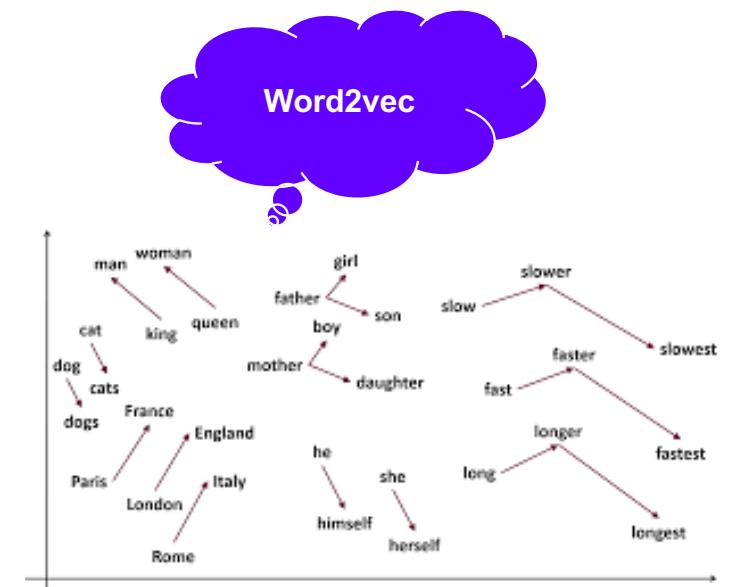
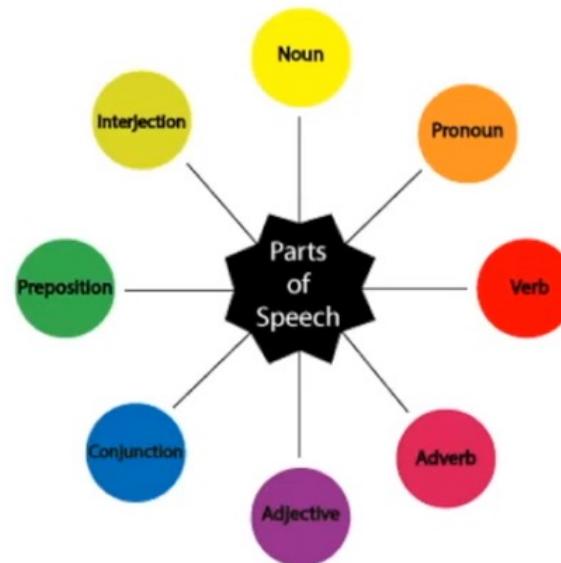
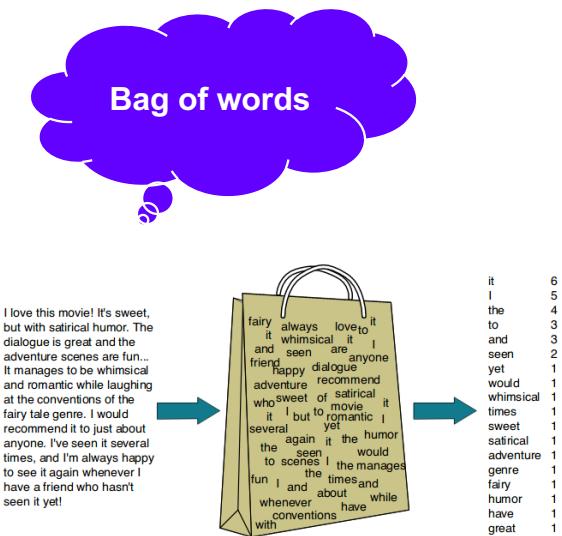
Computing is all you need...

The biggest lesson that can be read from 70 years of AI research is that general methods that leverage computation are ultimately the most effective, and by a large margin.

Rich Sutton, *The Bitter Lesson*



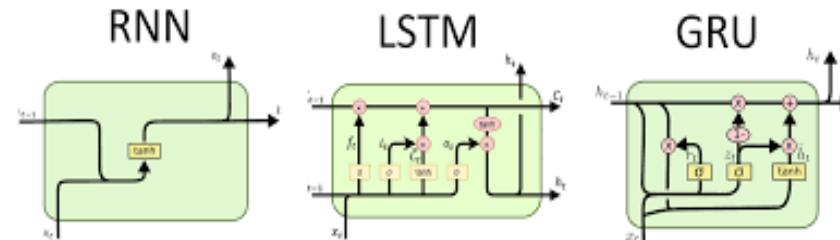
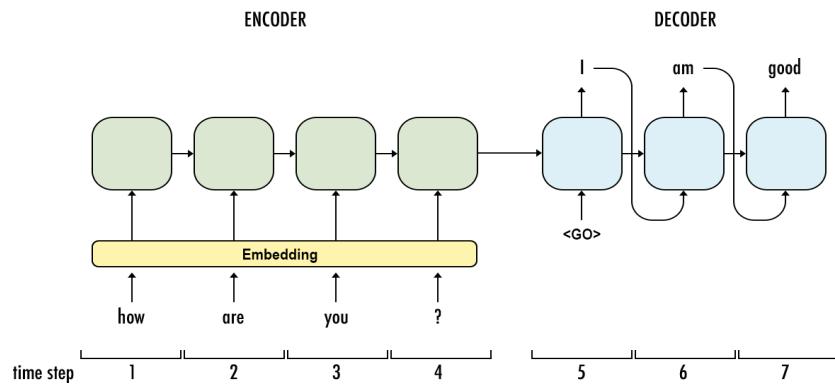
NLP In The Past



Most of natural language processing systems were based on simple statistical rules or non-complex Machine learning algorithms. The capabilities of these systems were limited to few tasks.

NLP In The Past Before 2017

Seq-to-Seq modeling



- Attention to the rescue
 - Cannot learn Long dependencies
 - Fails in Long sentences
- Recurrent
 - Sequential
 - Parallelization : not parallelizable

Emergence of new tasks with these new architectures :

- Translation
- Summarization
- Text completion
- ...

NLP Today

Attention Is All You Need, 2017

Google Brain, Google Research, and University of T

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
use@google.com

Llion Jones^{*}
Google Research
llion@google.com

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

Lukasz Kaiser^{*}
Google Brain
lukasz.kaiser@google.com

Illia Polosukhin^{*‡}
illia.polosukhin@gmail.com

Abstract

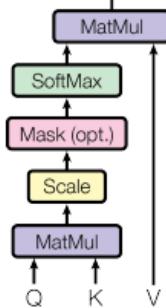
The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior to ones using recurrent neural networks, allowing for significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

1 Introduction

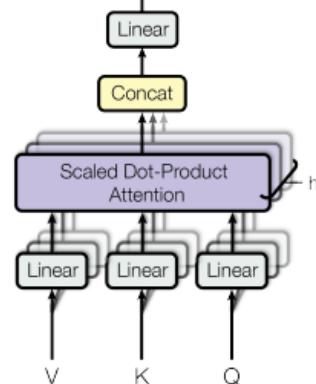


Attention mechanism

Scaled Dot-Product Attention



Multi-Head Attention

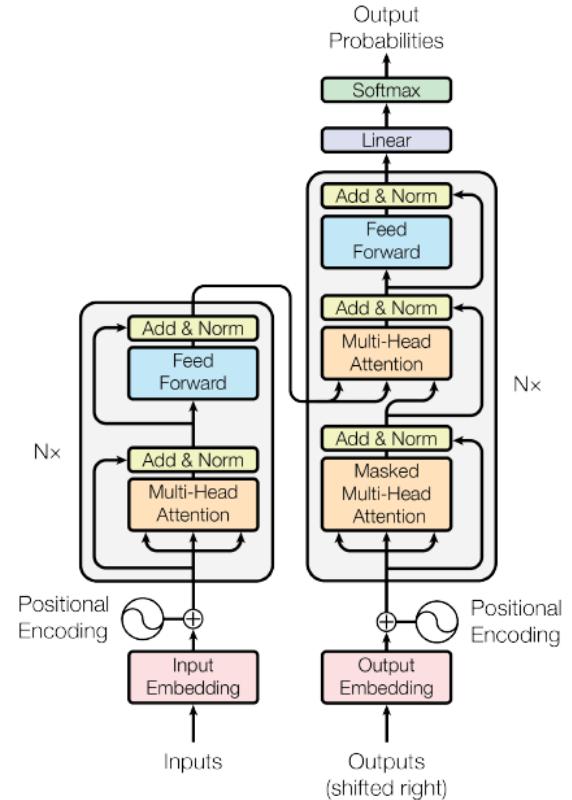


$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$



- Can learn Long dependencies
- Parallelizable

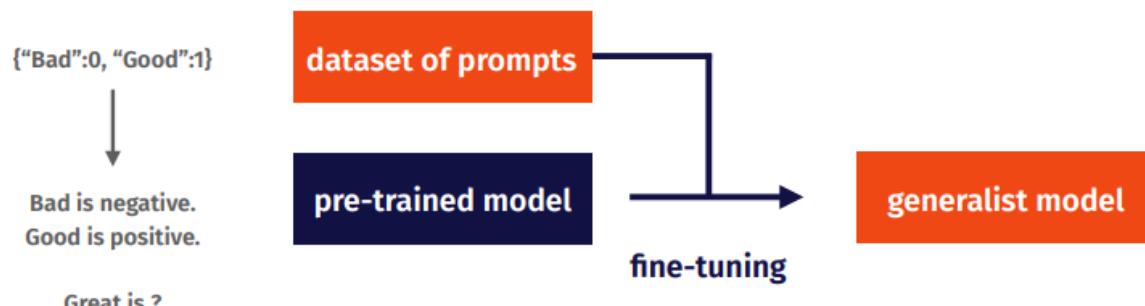
Transformers



LLM: Multi-Task Models At Scale, Unique Capabilities Arise Few-Shot Learning

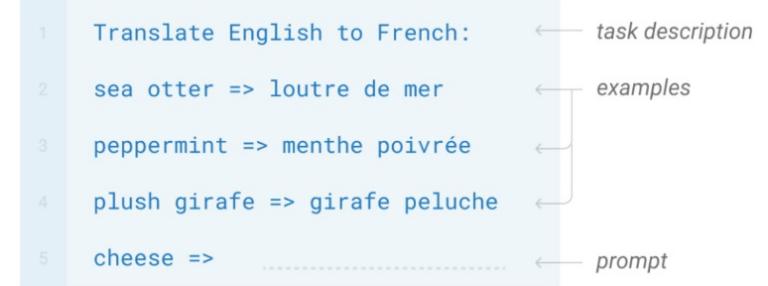
Power of extra-scale language models, **No need to fine-tuning** : Larger models can deal with unseen tasks on-the-fly

Language Models are Few-Shot Learners, Brown et al



Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

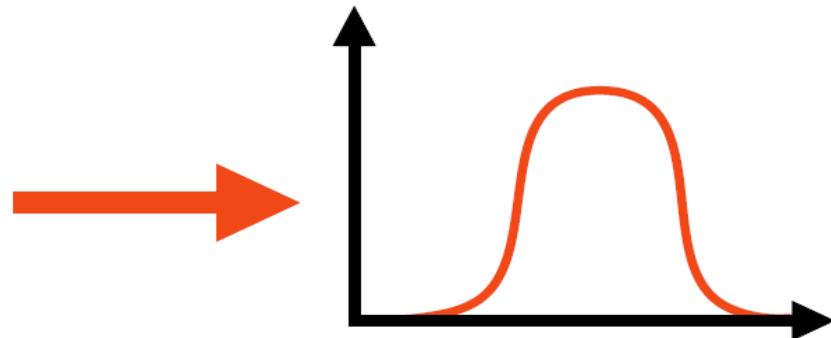


- Zero-shot : No example is provided, a description of the task only
- One-shot : One example is provided
- Few-shot : >1 examples are provided

What Does Modeling Language Really Mean?

Text corpus

...
In placerat nunc eu turpis pharetra, in imperdiet nunc porta. Etiam rhoncus odio ligula, ut pharetra
...
Integer egestas ullamcorper dapibus. Praesent at magna auctor, dignissim nulla eget, varius luctus.
Nam a molestie sem. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per
inceptos himenaeos. Integer sodales, eros ut congue sodales, dolor neque dignissim nunc, sit
amet mattis erat orci eu mettis. Nam viverra maximus cursus. Donec libero dolor, sollicitudin et
felis nec, mattis eleifend felis. Suspendisse vel erat maximus, viverra nisi eu, pellentesque urna.
Sed gravida nulla a odio aliquet, in commodo arcu ornare. Nam finibus quis magna et semper.
Etiam vel sem condimentum, tempor nibh a, cursus orci. Nulla eget aliquam neque. Suspendisse
potenti. Morbi maximus sem eu tellus viverra ultrices ac ut dolor. Cras luctus elit vel ligula
tincidunt, id varius metus rhoncus. Aliquam erat volutpat.



The capital of the United Arab Emirates is Abu **X**

$P(X | \text{context})$

Autoregressive/causal modelling

How Do We Learn/Model This Conditional Probability P?

🧠 Learning is easy: it's a **self-supervised** problem!

we can crawl vast amount of texts and simply predict the next word.

1st reason why extreme-scale
is possible!

🤔 We need a data structure to learn the distribution... **neural network?**

MLP/FC

✗ not sequence-aware

RNNs

✗ not scalable

SOTA until ~2018,
but slow & hard to train,
no long-term dependencies

Transformers

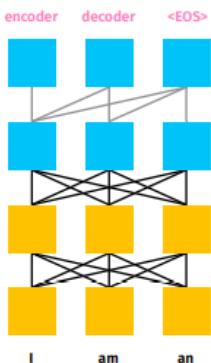


Natural Language Processing: The Age of Transformers Attention Mechanisms Revolutionized The Way We Do NLP

SOTA NLP models today are composed of a set of stacked transformers : Encoder-based, Decoder-based or Encoder-Decoder models.

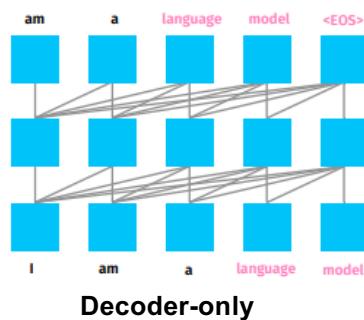
encoder-decoder

e.g. T5

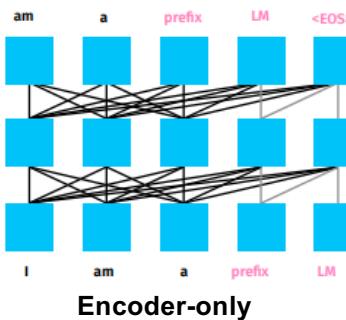


autoregressive LM

e.g. GPT



prefix LM



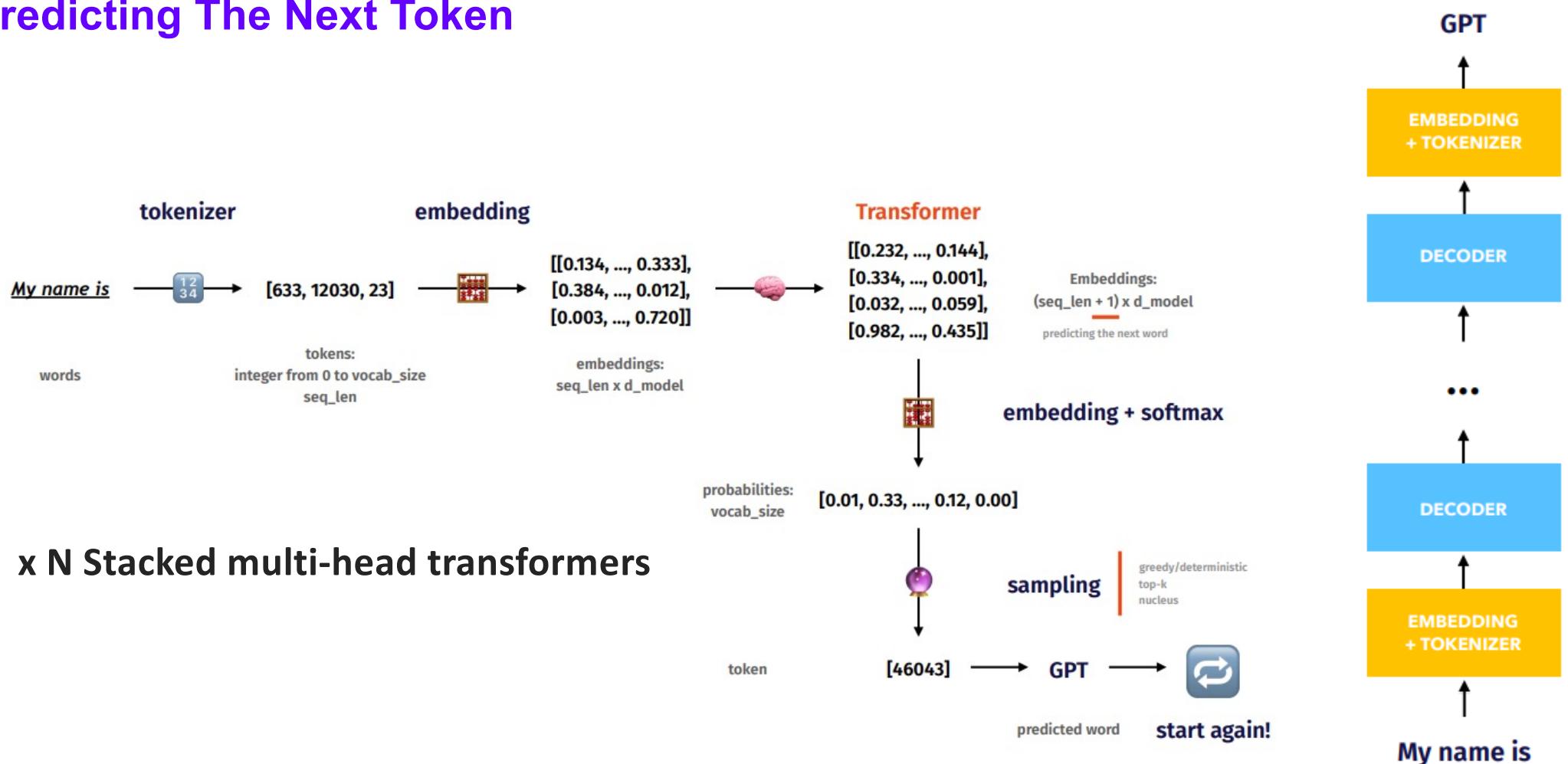
Examples of transformers-based language models :

- BERT
- ROBERTA
- GPT-2
- T5
- GPT-3
- PagNol
- Megatron-Turing NLG
- Noor

- Self-supervised models : MLM, next word, sentences order ...
- Parallelizable with teacher forcing

Focus on Decoder-only Architectures (E.g. GPT Models)

Predicting The Next Token



What LLM Can Do

Content creation		Content curation and analysis	Task automation
Text / Code	Idea generation	Rewrite / Summarize	Chatbot / Virtual assistant
	Text writing (book, training, course, plan...)	Text summary	Scheduling (meetings, tasks...)
Visuals / Sounds	Copywriting (emails, ads, blog posts...)	Video summary	Text editing (spell check, paraphrasing)
	Code writing (website, app...)	Audio transcript	Visuals editing (video cutting, image editing...)
	Image generation (text-to-image)	Language translation	Data cleansing
	Video generation (text-to-video)	Information clustering / formatting	Code auditing
	Voice generation (text-to-voice)	Information retrieval	Robot control
	Game design (AR, 3D design...)	Web search / Benchmarking	
		Q&A	
		Data analytics and forecast	
		Visual analytics	
		Sentiment / Intent recognition (ex. Fraud)	

Cost of the Models

Model	Training end	Chip type	TFLOP/s (max)	Chip count	Wall clock (days)	Total time (years)	Retail (US\$)	MMLU
GPT-3 175B	Apr/2020	V100	130	10,000	15 days	405y	\$9M	43.9
Llama 1 65B	Jan/2023	A100	312	2,048	21 days	118y	\$4M	63.4
Llama 2 70B	Jun/2023	A100	312	2,048	35 days	196y	\$7M	68.0
Titan 200B	Apr/2023	A100	312	13,760	48 days	1,319y	\$45M	70.4
GPT-4 1.7T	Aug/2022	A100	312	25,000	95 days	6,507y	\$224M	86.4
Gemini	Nov/2023	TPUv4	275	57,000	100 days	15,616y	\$440M	90.0
Llama 3 405B	Apr/2024	H100	989	24,576	50 days	3,366y	\$125M	85+
GPT-5	Apr/2024	H100	989	50,000	120 days	16,438y	\$612M	
Grok 2	Jun/2024	H100	989	20,000	50 days	6,571y	\$245M	
Olympus	Aug/2024	H100	989					
Gemini 2	Nov/2024	TPUv6	1,847					
Grok 3	Dec/2024	H100	989	100,000	50 days	32,855y	\$1.2B	

Alan D. Thompson. May/2024. LifeArchitect.ai

Table. Model training compute (see working, with sources⁸).

Choose your Weapon: Survival Strategies for Depressed AI Academics

Julian Togelius and Georgios N. Yannakakis*

April 14, 2023

Abstract

Are you an AI researcher at an academic institution? Are you anxious you are not coping with the current pace of AI advancements? Do you feel you have no (or very limited) access to the computational and human resources required for an AI research breakthrough? You are not alone; we feel the same way. A growing number of AI academics can no longer find the means and resources to compete at a global scale. This is a somewhat recent phenomenon, but an accelerating one, with private actors investing enormous compute resources into cutting edge AI research. Here, we discuss what you can do to stay competitive while remaining an academic. We also briefly discuss what universities and the private sector could do improve the situation, if they are so inclined. This is not an exhaustive list of strategies, and you may not agree with all of them, but it serves to start a discussion.

1 Introduction

As someone who does AI research in a university, you develop a complicated relationship to the corporate AI research powerhouses, such as DeepMind, Open AI, Google Brain and Meta AI. Whenever you see one of these papers that train some kind of gigantic neural net model to do something you weren't

Talk Outline

- A General Introduction
- Our Work on Generative AI

Noor Released in 2021

Technology Innovation Institute Announces Launch of NOOR, the World's Largest Arabic NLP Model

13 Apr, 2022



Model is most powerful one in Arabic language to date with 10 billion parameters

Features applications in automated summarization, chatbots, personalized marketing

Mon 11-04-2022 15:05 PM

Technology Innovation Institute announces launch of NOOR, world's largest Arabic NLP model



A- A A+

ABU DHABI, 11th April, 2022 (WAM) -- Technology Innovation Institute (TII), a global research centre and applied research pillar of Abu Dhabi's Advanced Technology Research Council (ATRC), today announced the launch of NOOR, the world's largest Arabic natural language processing (NLP) model to date.

TII's team of advanced researchers and Artificial Intelligence (AI) specialists at its AI Cross-Centre Unit, joined forces on this initiative with LightOn, a technology company that unlocks extreme-scale machine intelligence for businesses, to revolutionise Arabic NLP models.

The NOOR model carries out varied, cross-domain tasks simply from natural language instructions. Its training dataset is the world's largest high-quality Arabic dataset, combining web data with books, poetry, news articles, and technical information to significantly widen the applicability of the model.

Dr. Ray O. Johnson, CEO, TII and ASPIRE, said, "With this development, we are on track to boost our research capabilities and credentials in AI, as well as elevating the status of Abu Dhabi and the UAE as a serious research ecosystem. Our expert teams have demonstrated yet again that this region can achieve breakthrough R&D outcomes that impact the world."

Dr. Ebtesam Almazrouei, Director, AI Cross-Centre Unit, TII, commented, "Large language models have taken the world of natural language processing by storm, and we are proud to introduce this cutting-edge model with 10 billion parameters, the world's largest Arabic NLP model. The uniquely large Arabic dataset collected to train the model is the result of months of work that included curating, scrapping, and filtering of varied sources."

Speaking on the upcoming launch, Prof. Mérourane Debbah, Chief Researcher, Digital Science Research Centre and AI Cross-Centre Unit, TII, said, "With NOOR, TII has expanded the scope of the modern standard Arabic model by leveraging know-how in large language models to build cross-disciplinary, cutting-edge expertise in this new generation of AI research. This is only the first step in TII's efforts to contribute to the wider UAE Strategy for Artificial Intelligence, through supporting AI integration across key sectors of the economy."

The model is named for the Arabic word "light", and it represents the UAE's global contribution to advanced technology and artificial intelligence.

WAM/Tariq alfaham/Khoder Nashar

Noor

Released April, 2022

Arabic language model with
10B parameters

First published Arabic Article by Noor November, 2022

Noor Arabic language
model generates an article

The screenshot shows a news article from the Global Media Congress website. The headline reads: "أهم مذكرة صحفية عن الاعدام الوطني في مصر الرسمية تناول «الكونغرس» المنشاوي". Below the headline, there is a large image of a robotic hand holding a red pen, with the Noor logo overlaid. The main text of the article discusses the publication of Noor's first Arabic article, which is described as a breakthrough in AI-generated journalism. The article highlights the use of Noor's 10B parameter Arabic language model to generate the text. It also mentions that Noor has been trained on 100,000 Arabic news articles and can now write news stories in Arabic. The author of the article is Dr. Abeer Almousawi.

الأخاد

GLOBAL MEDIA CONGRESS

«الأخاد» تنشر أول مقال مكتوب بـ تقنيات الذكاء الاصطناعي

حل يمكننا إيقاد كوكب الأرض؟

النمو الاقتصادي وحده لن يحل مشكلة تغير المناخ

علماء يكشفون عن نتائج دراسة مبتكرة

النحو العربي يفتح آفاقاً جديدة في تطوير الذكاء الاصطناعي

الذكاء الاصطناعي والرسالة الإعلامية

Noor - Narrator

The screenshot shows a web browser window for the 'Noor' application, specifically the 'Narrator' section. The title bar says 'Noor' and the address bar shows 'localhost:3000/#/narrator'. The main interface has a green header with the Arabic word 'راوي' (Narrator) and a back button. Below the header is a large input field for text entry, with a note above it: 'الرجاء إضافة نص ينكون من 5 حرف على الأقل'. Inside the input field, the text 'كرة القدم رياضة شعبية' is typed. A note below the input field says '21 حرف'. At the bottom of the input field is a green button labeled 'إنشاء نص' (Create Text). To the right of the input field, there is a small circular icon with a pen symbol. At the bottom right of the page, there is a note: 'النص المنتج' and 'ترقب التلخيص هنا' with a downward arrow icon. The footer contains the text 'Made with ❤️ At Technology Innovation Institute'.

Noor - Summarization

The screenshot shows a web-based summarization tool. At the top, there's a green header bar with a back button and a 'Summarization' button. Below the header is a toolbar with icons for file operations like 'New', 'Open', 'Save', etc. The main content area has a title 'تلخيص النص' (Summary) and a sub-instruction 'الرجاء إضافة نص ينكر من 500 حرف على الأقل' (Please add a text to summarize from at least 500 characters). The text area contains a summary of a historical document about the formation of the UAE. A green button at the bottom says 'لخص النص' (Summarize text). On the right side, there's a sidebar with the text 'النص المنتج' (Generated text) and a link 'ترقب التلخيص من هنا' (Check the summary here).

الرجاء إضافة نص ينكر من 500 حرف على الأقل

العودة إلى الصفحة الرئيسية

تلخيص النص

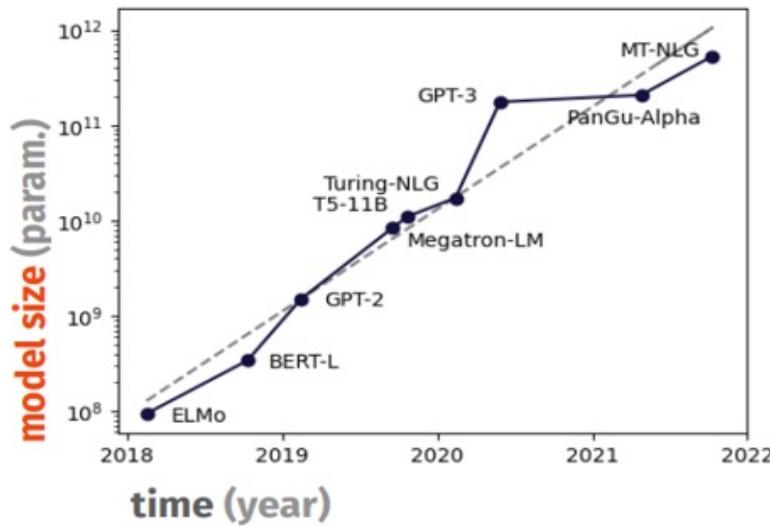
يرتبط تاريخ دولة الإمارات العربية المتحدة المعاصرة بتاريخ "ساحل عمان". والذي يضم في يومنا هذا سلطنة عمان ودولة الإمارات العربية المتحدة. وفي القرن السابع عشر وما بعده أصبح الخليج العربي محطة توقف إستراتيجية مهمة لسفن التجار المتسافرين بين أوروبا وأسيا. وفي عام 1892، انضمت الإمارات التي كانت تعرف حتى ذلك الوقت بـ"إمارات الساحل المتصالح" معاً لتوقيع معاهدة مع بريطانيا العظمى، تقوم بموجبها بريطانيا العظمى بحماية السواحل البحرية ضد أي عدوان بحري وتتوفر الدعم والمساعدة ضد أي عدوان بري لهذه الإمارات. وفيما بعد خلال العام 1968، أعلنت بريطانيا أنها ستنهي علاقتها التغاذية مع إمارات الساحل المتصالح وقطر والبحرين. واستنسحب من منطقة الخليج العربي. وبمبادرة من المغفور له الشيخ زايد بن سلطان آل نهيان، تم تأسيس دولة الإمارات العربية المتحدة في الثاني من ديسمبر 1971. التاريخ الذي تختلف به الدولة كل عام بعيداً عنها. كدولة اتحادية مكونة من سنت إمارات، هي: أبوظبي، دبي، والشارقة، وعجمان، وأم القيوين، والفجيرة، وفي العام التالي، انضمت إمارة رأس الخيمة للاتحاد.

لخص النص

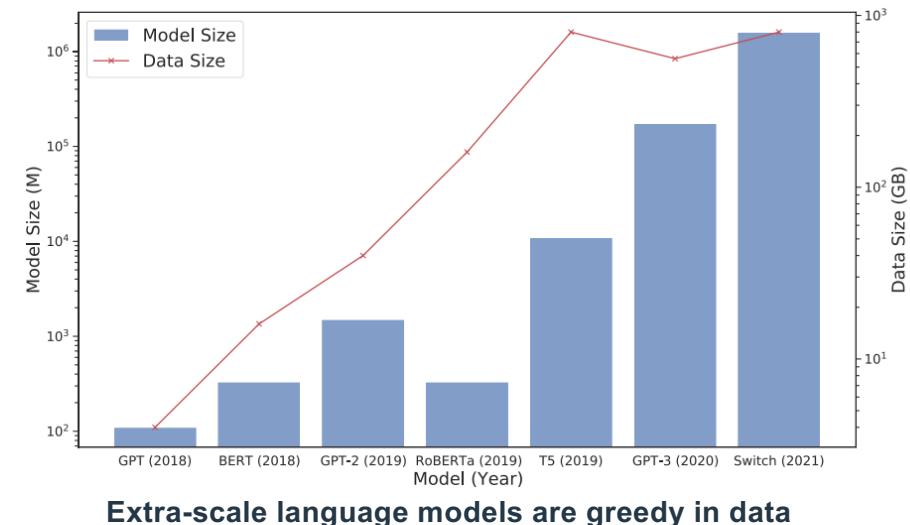
النص المنتج

ترقب التلخيص من هنا

Extreme-scale Models for NLP Go Big to Boost The Performance



Over the last four years, the size of state-of-the-art language models has doubled every 3-4 months



Extra-scale language models are greedy in data

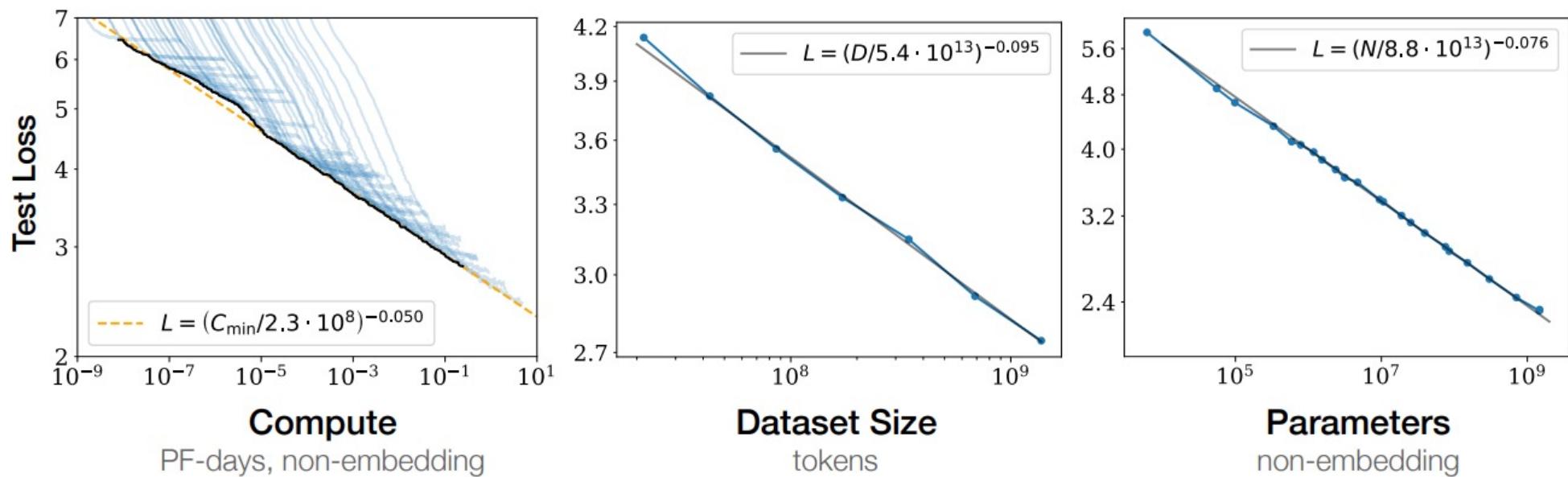
Model	Size (# Parameters)	Training Tokens
LaMDA (Thoppilan et al., 2022)	137 Billion	168 Billion
GPT-3 (Brown et al., 2020)	175 Billion	300 Billion
Jurassic (Lieber et al., 2021)	178 Billion	300 Billion
Gopher (Rae et al., 2021)	280 Billion	300 Billion
MT-NLG 530B (Smith et al., 2022)	530 Billion	270 Billion

Most models are trained for approximately 300 billion tokens

Performance depends strongly on scale, weakly on model shape

Extreme-scale Language Models Scaling Laws

Kaplan et al. : Language modeling performance improves smoothly as we increase the model size, dataset size, and amount of compute used for training. For optimal performance, all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.

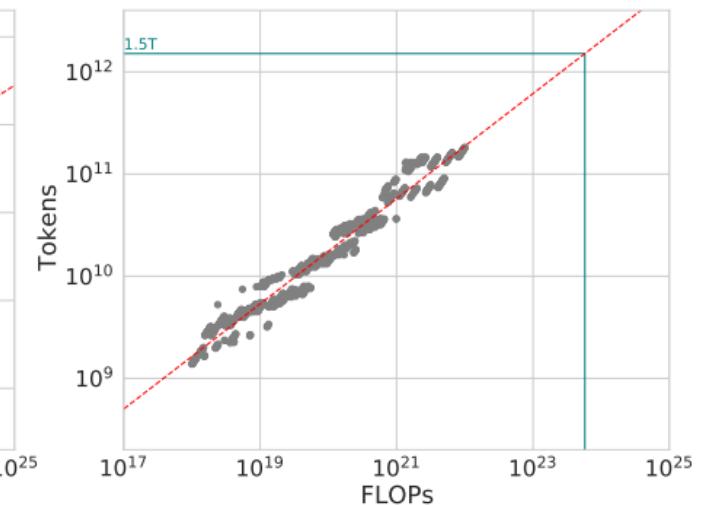
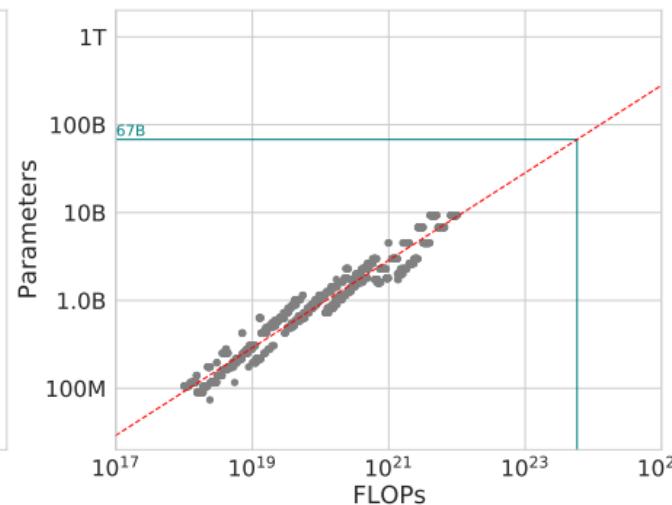
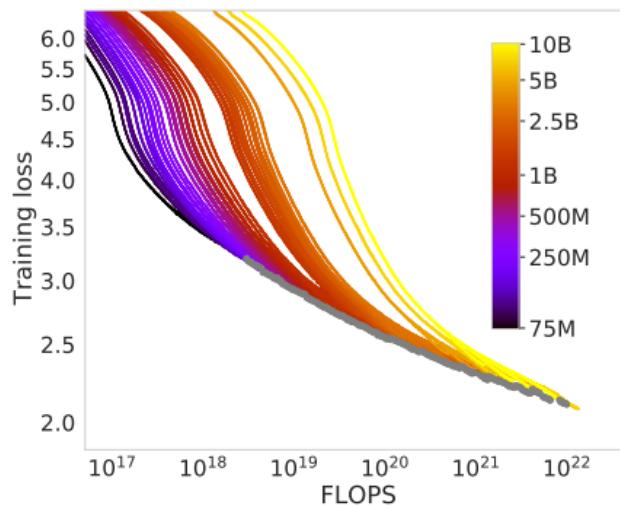


Trends In Model Scaling

New Paper of DeepMind Revisited Kaplan et al.

Current large language models are significantly undertrained, a consequence of the recent focus on scaling language models whilst keeping the amount of training data constant.

Under a **compute budget constraint**, one should determine the corresponding number of parameters and tokens to achieve the best possible loss.



Example. Chinchilla uses the same compute budget as Gopher (280B) but with 70B parameters and 4x more data. It uniformly and significantly outperforms Gopher (280B), GPT-3 (175B), Jurassic-1 (178B), and Megatron-Turing NLG (530B) on a large range of downstream evaluation tasks.

Training Compute-Optimal Large Language Models

Jordan Hoffmann*, Sebastian Borgeaud*, Arthur Mensch*, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Webl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals and Laurent Sifre*

*Equal contributions

We investigate the optimal model size and number of tokens for training a transformer language model under a given compute budget. We find that current large language models are significantly under-trained, a consequence of the recent focus on scaling language models whilst keeping the amount of training data constant. By training over 400 language models ranging from 70 million to over 16 billion parameters on 5 to 500 billion tokens, we find that for compute-optimal training, the model size and the number of training tokens should be scaled equally: for every doubling of model size the number of training tokens should also be doubled. We test this hypothesis by training a predicted compute-

The Falcon Series of Open Language Models: 7B, 40B, 180B .

One of the World's Most Powerful Open Access Large Language Model

- Falcon LLM is currently at the **top of the Hugging Face Leaderboard for pre-trained Open Large Language Models.**
- Falcon-180B, has been trained on over 3.5 trillion tokens of text—the **largest openly documented pretraining run.**
- Falcon-180B** significantly outperforms models such as PaLM or Chinchilla, and improves upon concurrently developed models such as LLaMA 2 or Inflection-1.
- The production training for Falcon 180B was run on 4096 GPUs using the Amazon SageMaker cloud machine learning platform for a total of about 7,000,000 GPU hours.

	Falcon-7B	Falcon-40B	Falcon-180B
Pretraining [tokens]	1,500B	1,000B	3,500B
Compute [PF-days]	730	2,800	43,500
Training [A100s]	384	384	4,096
Availability	Apache 2.0	Apache 2.0	Responsible use license
Agg. performance (60.8	67.1	70.3
Closet model	<GPT-3	Chinchilla	PaLM-2 Large



Some Data ingredients that have worked for us

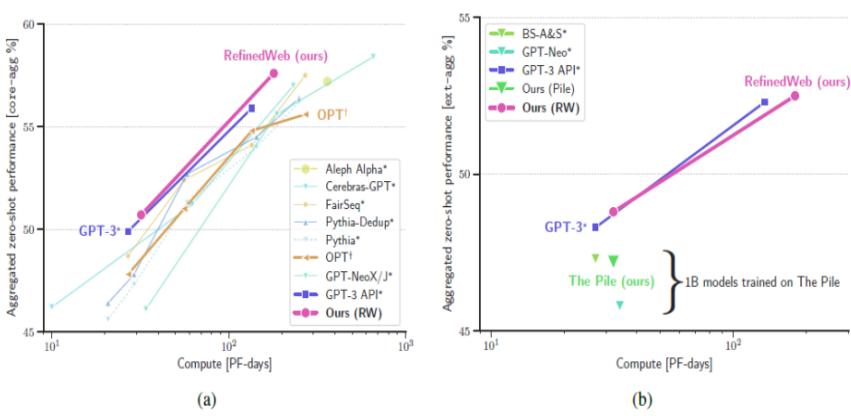


Figure 3. Models trained on **REFINEDWEB** alone outperform models trained on curated corpora. Zero-shot performance averaged on our core-agg (left) and ext-agg (right) task aggregates (see Section 4.1 for details, and Figure 1 for results on main-agg). Existing open models fail to match the performance of the original GPT-3 series (left); however, models trained on RefinedWeb significantly outperform models trained on **The Pile**: including our direct comparison model (right), ruling out our pretraining setup as the main source of increased performance. In fact, our RefinedWeb models even match the performance of the **GPT-3** models.

Table 1. **REFINEDWEB** improves on existing English pretraining datasets for large language models by combining extensive filtering with stringent deduplication at unprecedented scale. For additional details, see the full version in Table 12 of Appendix F.3.

Dataset	Size	Availability	Web	CC Processing	Deduplication
MASSIVE WEB DATASETS					
C4	~ 360GT	Public	100%	Rules + NSFW words blocklist	Exact: spans of 3 sentences
OSCAR-21.09	~ 370GT	Public	100%	Built at the line-level	Exact: per line (~ 55% removed)
OSCAR-22.01	~ 283GT	Public	100%	Line-level rules + optional rules & NSFW URL blocklist	Exact: per line (optional, not used for results in this paper)
CURATED DATASETS					
■ GPT-3	300GT	Private	60%	Content filter trained on known high-quality sources	Fuzzy: MinHash (~ 10% removed)
▼ The Pile	~ 340GT	Public	18%	jusText for extraction, content filter trained on curated data	Fuzzy: MinHash (~ 26% removed)
★ PaLM	780GT	Private	27%	Filter trained on HQ data	Unknown
OURS					
● RefinedWeb	~ 5,000GT	Public (600GT)	100%	trafilatura for text extraction, document and line-level rules, NSFW URL blocklist	Exact & fuzzy: exact substring+MinHash (~ 50% removed)

Is is the data **problem** an easy one?

🚀 We only need **unsupervised** data: it's plenty available.

Common Crawl



CommonCrawl: public web data,
200TB/2-3 months.

🤔 But what about **data quality**?

The “**Google approach**”: trillions of tokens, no filtering (e.g. C4)

The “**OpenAI/EAI approach**”: curated dataset, filtering for CC data (e.g. The Pile)
and others!

➡ This also depends on the **type of model** being built: more important for **generative models!**

e.g. GPT

Two quality problems: data **content** & **complexity**

💡 **Generative models will reflect content of the data: pay attention to dataset content!**

for models dedicated to classification/embeddings, less important.

💡 **Zero-shot performance will depend on data quality:**

long-term correlations books, technical papers, etc.

quality writing books, quality websites, etc.

soft prompting explanatory books, exercise websites, etc.

special vocabulary technical papers, code, etc.

Examples of dataset composition

🧠 OpenAI GPT-3 training data:

300B
tokens

41%
curated data

3%
wikipedia

22%
curated web

16%
books

OpenAI GPT-3 dataset

	Dataset size [GT]	Epochs [/]	Tokens seen [GT]	Weight [%]
WebText2	19		3,47	0,22
Books1	12		2,00	0,08
Books2	55		0,44	0,08
Wikipedia	3		3,00	0,03
Curated	89		123	0,41
CommonCrawl	410		0,44	0,6
Bulk	410		180	0,6
Training set		300		

high-quality data
is repeated

Examples of dataset composition

 EleutherAI The Pile:

1.2TB

data

82%

curated data

1.5%

wikipedia

15%

curated web

16%

books

Component	Raw Size	Weight	Epochs	Effective Size	Mean Document Size
Pile-CC	227.12 GiB	18.11%	1.0	227.12 GiB	4.33 KiB
PubMed Central	90.27 GiB	14.40%	2.0	180.55 GiB	30.55 KiB
Books3 [†]	100.96 GiB	12.07%	1.5	151.44 GiB	538.36 KiB
OpenWebText2	62.77 GiB	10.01%	2.0	125.54 GiB	3.85 KiB
ArXiv	56.21 GiB	8.96%	2.0	112.42 GiB	46.61 KiB
Github	95.16 GiB	7.59%	1.0	95.16 GiB	5.25 KiB
FreeLaw	51.15 GiB	6.12%	1.5	76.73 GiB	15.06 KiB
Stack Exchange	32.20 GiB	5.13%	2.0	64.39 GiB	2.16 KiB
USPTO Backgrounds	22.90 GiB	3.65%	2.0	45.81 GiB	4.08 KiB
PubMed Abstracts	19.26 GiB	3.07%	2.0	38.53 GiB	1.30 KiB
Gutenberg (PG-19) [†]	10.88 GiB	2.17%	2.5	27.19 GiB	398.73 KiB
OpenSubtitles [†]	12.98 GiB	1.55%	1.5	19.47 GiB	30.48 KiB
Wikipedia (en) [†]	6.38 GiB	1.53%	3.0	19.13 GiB	1.11 KiB
DM Mathematics [†]	7.75 GiB	1.24%	2.0	15.49 GiB	8.00 KiB
Ubuntu IRC	5.52 GiB	0.88%	2.0	11.03 GiB	545.48 KiB
BookCorpus2	6.30 GiB	0.75%	1.5	9.45 GiB	369.87 KiB
EuroParl [†]	4.59 GiB	0.73%	2.0	9.17 GiB	68.87 KiB
HackerNews	3.90 GiB	0.62%	2.0	7.80 GiB	4.92 KiB
YoutubeSubtitles	3.73 GiB	0.60%	2.0	7.47 GiB	22.55 KiB
PhilPapers	2.38 GiB	0.38%	2.0	4.76 GiB	73.37 KiB
NIH ExPorter	1.89 GiB	0.30%	2.0	3.79 GiB	2.11 KiB
Enron Emails [†]	0.88 GiB	0.14%	2.0	1.76 GiB	1.78 KiB
The Pile	825.18 GiB			1254.20 GiB	5.91 KiB

compared to OpenAI, extra modalities:

42%

technical data

9.1%

code & maths

8.8%

conversational

GPT-J outperforms GPT-3 on code...

but underperforms on aggregate bench.

Some good sources of data



Web data: [CommonCrawl](#).

Common Crawl



+ filtering solution: [CCNet](#)

[Curated crawls](#) from link repositories. [Reddit](#), [HackerNews](#), etc.



Books: “open-source” books from [Gutenberg](#).

other, more or less legal sources.

Gears icon: Technical: [wikipedia](#), paper repositories ([arXiv](#), [HAL](#)), patent offices, etc.

Speech bubble icon: Conversational: subtitles, podcasts, forums, etc.

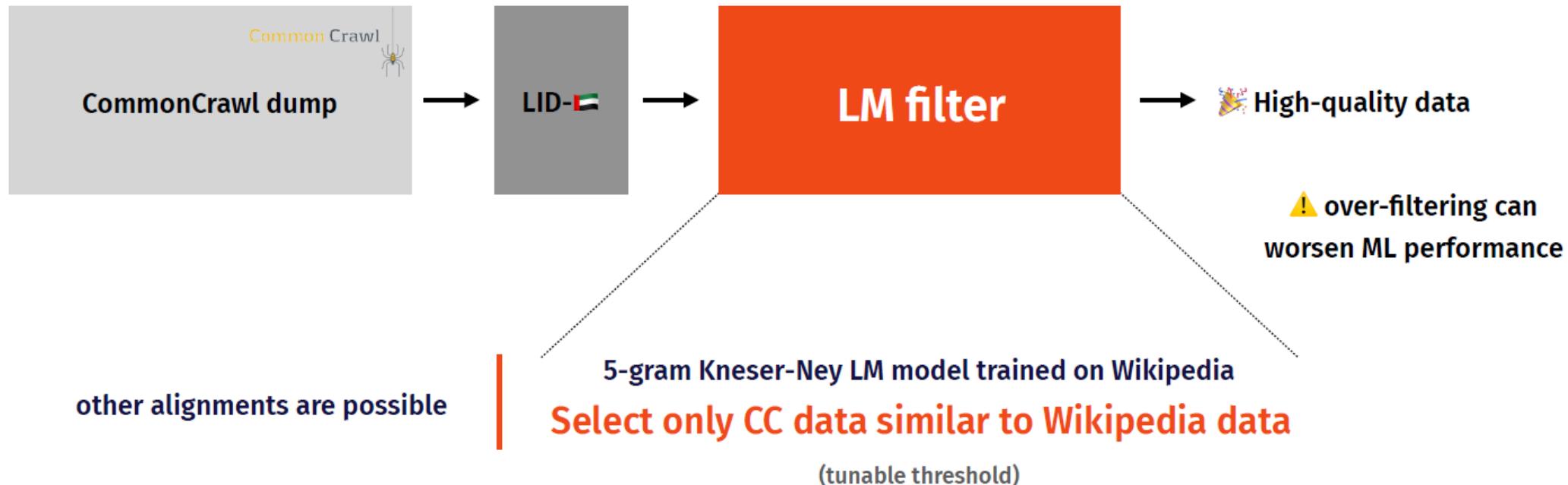
Government building icon: Governmental open-data.

Making CommonCrawl more palatable

⚠ The problem: raw CommonCrawl data is **really, really bad!**

spam, ads, NSFW, etc.

👮 CCNet: Extracting **High Quality** Monolingual Datasets from Web Crawl Data, Wenzek et al.



How much data do you **actually** need?

✓ Can be determined through **scaling laws**: data needs don't scale exponentially with size 😊.

GPT-3 frontier			
	model size	optimality	convergence
GPT-2	1.3B	23	156
	2.7B	34	228
GPT-J	6.7B	54	366
	13B	77	517
GPT-3	50B	155	1041
	175B	300	2000
MT-NG	530B	528	3554
	1T	735	4944

actually see improvements
even up to 300B!

start of good zero-shot

Other aspects: **tokenization**, **anonymisation**, etc.

 **Tokenization:** no smart solution, usually byte-pair encoding with SentencePiece.

train on ~1% of final dataset, enable fallback byte tokens.

 **Anonymisation:** some interesting ideas around removing phone numbers, e-mails, etc.
with dedicated tokens.

 **Content blocklists?** Better to implement in inference pipeline.

detoxification, etc. rarely works in practice...

 **Legal aspects:** complicated, and jurisdiction-dependent. Some backing for crawling...

Some Training ingredients that have worked for us

3D parallelism



Data, pipeline, tensor parallel combined.

ZeRO-1



Shard optimizer states.

Increased complexity over FSDP, but also increased **flexibility** for low-resource infrastructure.

FlashAttention



Optimized attention implementation.

Triton



Custom kernels for operations.

"Monolayer"



180B@4,096 A100s:
<15GB memory use

Fusing operations is the main driver of increased throughput.

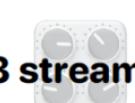
Modeling



Limited modeling interventions,
mostly for **inference scalability**.

Multigroup attention,
parallel attention/FFN

S3 streaming



No distributed filesystem,
dataformat designed for S3.

Simplify distributed infrastructure.

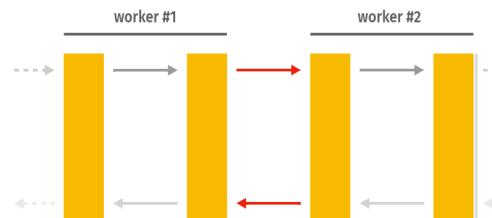
Large and Deep Transformer-based Neural Architectures

How Is It Possible to Train in Reasonable Time? Parallelism in The Scope

- Data Parallelism : Partition Mini-batches over multiple workers with copies of the networks



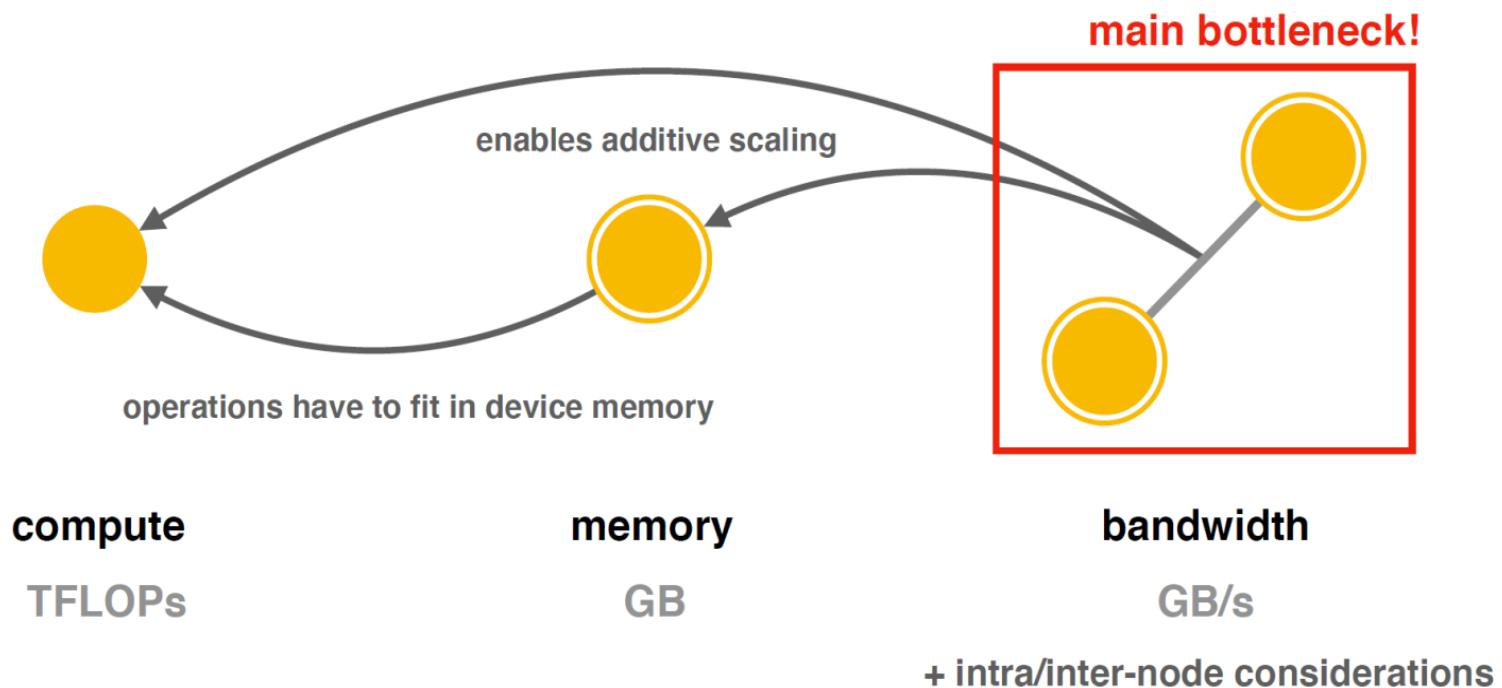
- Pipeline Parallelism : Horizontal parallelization over the layers



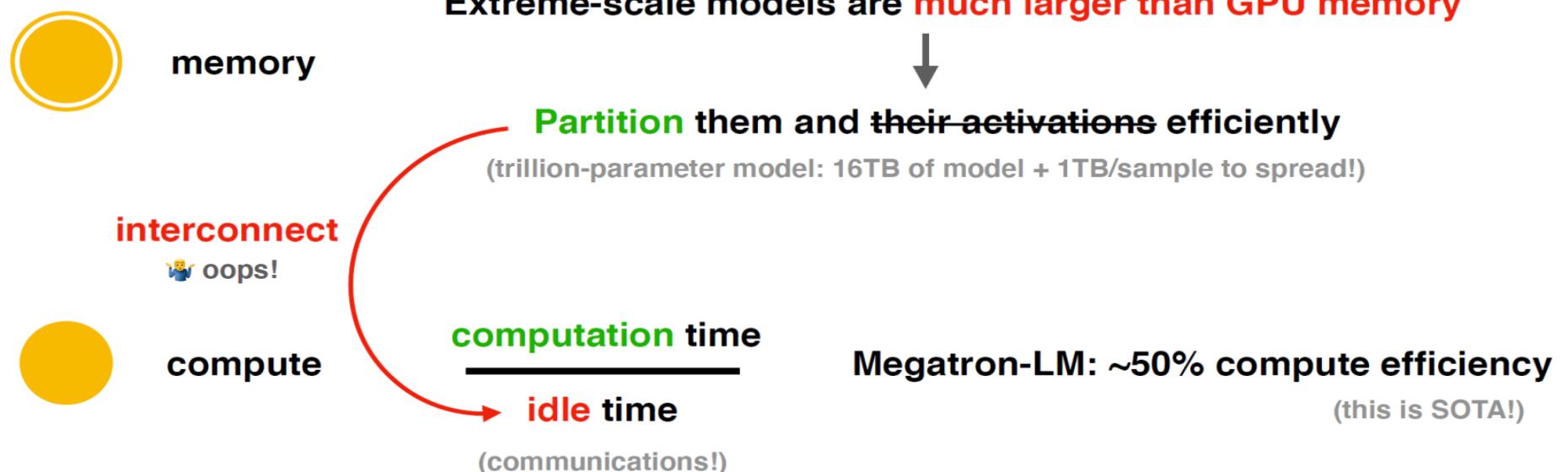
- Tensor Parallelism : Parallelism at the layer level



It's all about tradeoffs: compute, memory, bandwidth



Two key challenges: compute and memory efficiency

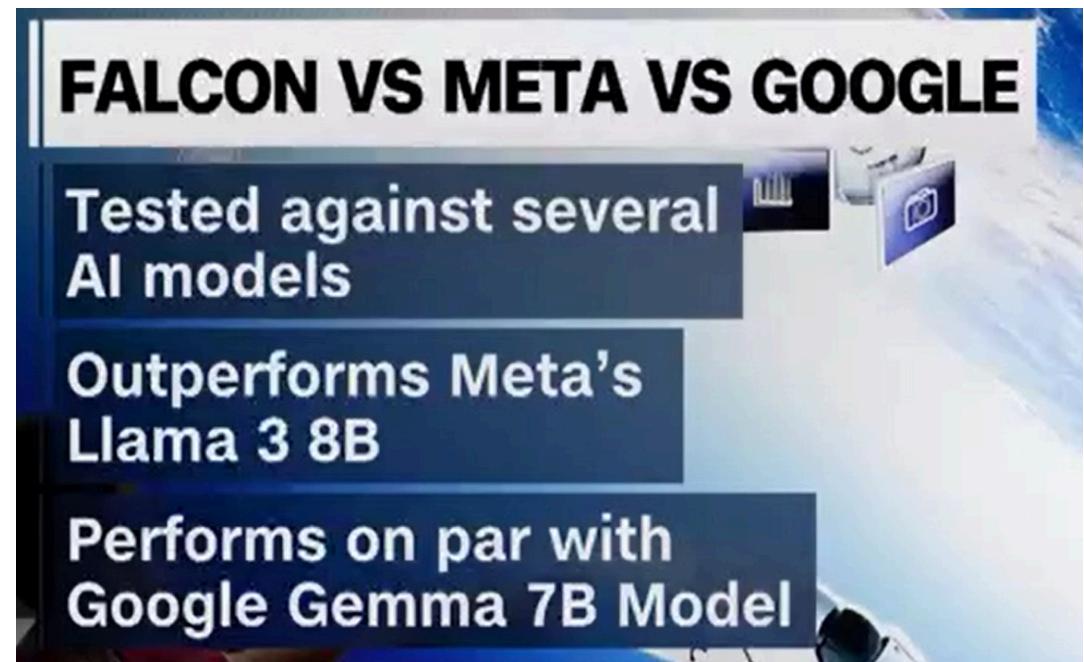


At 50% efficiency, 4,000 A100s train a trillion-parameter model in 100 days.

Falcon 2: 13th May 2024, top ranked Open Source Model

Falcon 2: UAE's Technology Innovation Institute Releases New AI Model Series, Outperforming Meta's New Llama 3

- Next-Gen Falcon 2 Series launches AI Model that is Open-Source, Multilingual, and Multimodal – and is only AI Model with Vision-to-Language Capabilities
- New Falcon 2 11B Outperforms Meta's Llama 3 8B, and Performs on par with leading Google Gemma 7B Model, as Independently Verified by Hugging Face Leaderboard
- Immediate Plans Include Exploring 'Mixture of Experts' for Enhanced Machine Learning Capabilities

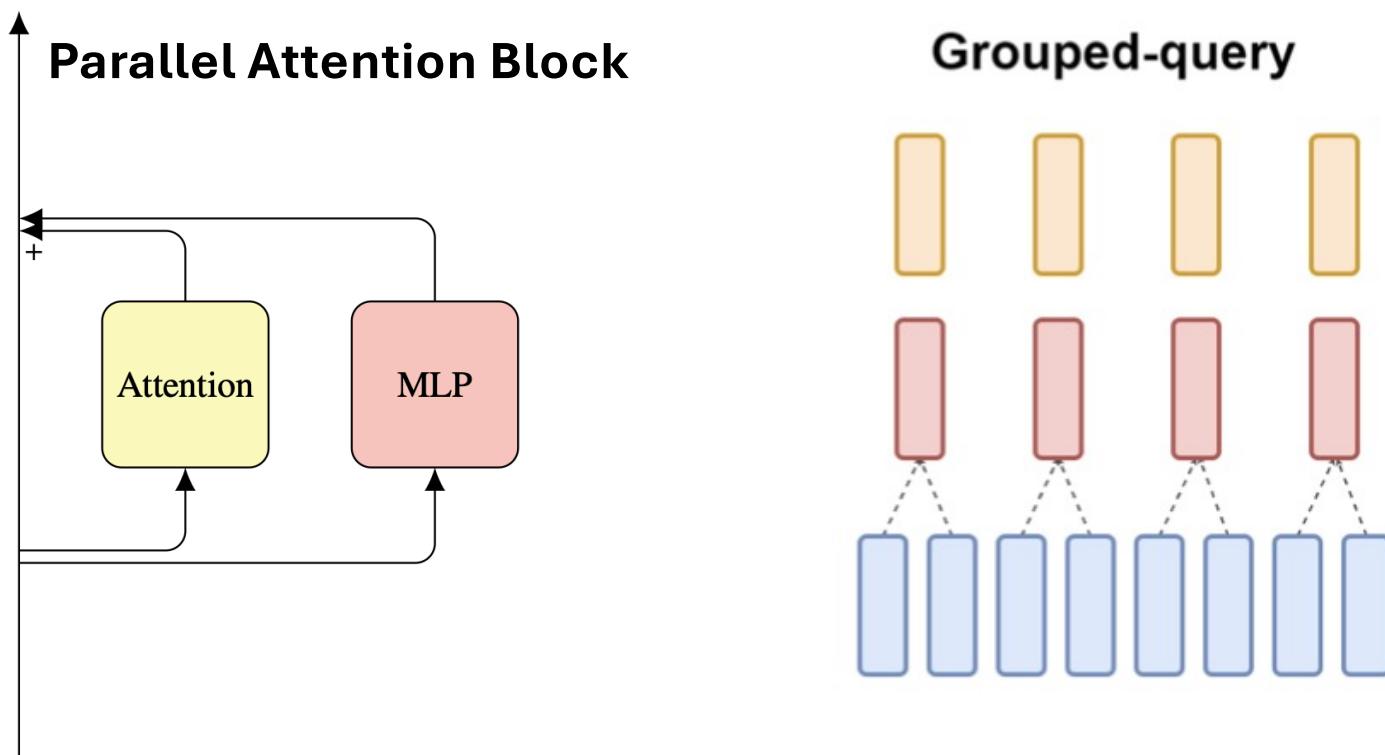


Model Architecture and Hyperparameters

Vocabulary size	65024
Activations	GELU
Sequence length	2048 to 8192
Layers	60
Embeddings dimension	4096
Head dimension	128
Num. Query Heads	32
Num. Key/Value Heads	8
Feed forward factor	4

Precision	BF16
Optimizer	AdamW
	Beta1 = 0.99
	Beta2 = 0.999
	Eps = 1e-7
Learning Rate	Linear from 0 to 3.7e-4 over 4GT
	Cosine decay to 1.89e-5 up to 4500 GT, constant afterwards
Weight decay	1e-1
Z-loss	1e-4
Batchsize	Variable

Model Architecture and Hyperparameters



Training Data

Training stage	Context Length	Billion Tokens
Stage 1	2048	4500
Stage 2	4096	250
Stage 3	8192	250
Stage 4	8192	500

Data mixture

RefinedWeb (English web data)

RefinedWeb-Europe (10 languages)

High quality technical data

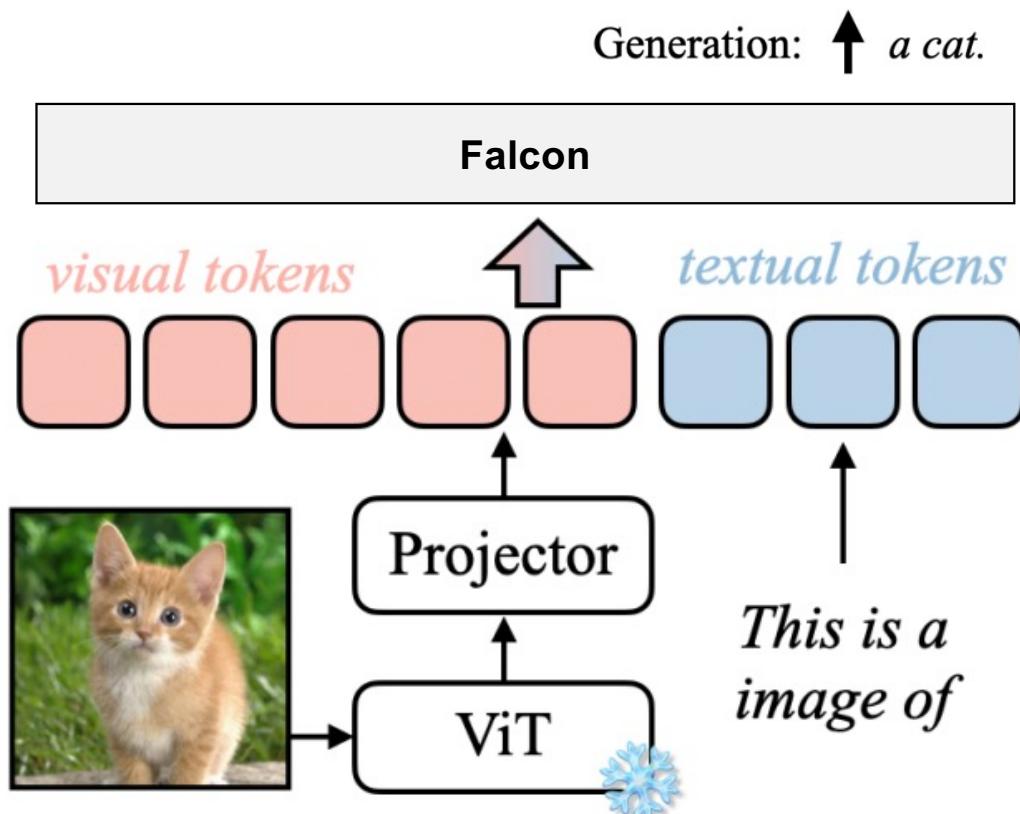
Code (43 programming languages)

Conversational data

10B Stage 1 Relative Multilingual Da

cs	de	es	fr	it	nl	pl	pt	ru	zh
3.63	20.97	20.58	16.59	10.45	6.28	6.93	5.51	2.02	0.07

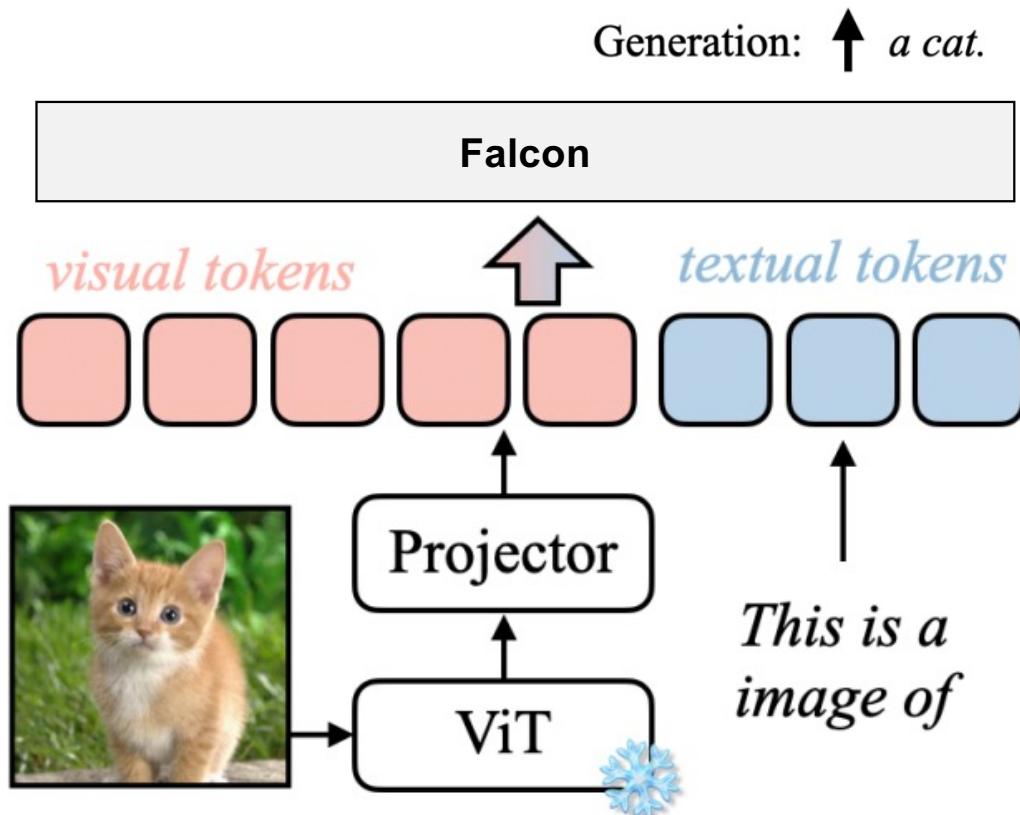
Extending Falcon2-11B to handle images



What's different from LLMs?

- Text is easier to tokenize as it's discrete.
- Images require a deep learning model to tokenize them, so called soft tokens.
- Not any image encoder, but one aligned with the text space such as CLIP.

Training Procedure



Training stages:

- **Stage 01:** Projector initialization. The LLM and ViT are separately pre-trained, while the projector is usually initialized from random weights. Therefore, we first pre-train the projector while freezing both ViT and LLMs on image-caption pairs.
- Stage 02: We then further perform instruction tuning of the LLM and projector on image instruction datasets

Examples: History

Falcon-VL for Image understanding

Upload an Image and Enter a Prompt to Generate a Caption using Falcon-10b-VL

image

A NORIA, OR WATER WHEEL, BUILT CIRCA 450 C.E., HAMA, SYRIA



The image shows a historical water wheel system, known as a noria, located in Hama, Syria. It consists of several large, circular stone structures mounted on a central vertical axis, used to lift water from a river or canal. The surrounding area appears to be a mix of ancient ruins and modern infrastructure.

prompt

the image best illustrates which continuities in world history?

Clear

Submit

output

The image best illustrates the continuity in world history of humans adapting to their environments. The scene features a large, old stone building with a water wheel, which is a structure that has been used for centuries to harness the power of water to perform tasks such as grinding grain or pumping water. This demonstrates how humans have utilized natural resources like water to develop technologies and infrastructure that have allowed them to thrive in various environments. The water wheel in the image serves as a reminder of the ingenuity and resourcefulness of humans throughout history, as they have adapted to their surroundings and utilized available resources to meet their needs.

Flag

TelecomGPT: A Step forward to Self Driven Autonomous Networks

UAE's TelecomGPT: The AI Breakthrough Set to Transform Telecoms

17 JULY 2024



TELECOM Review
THE TELECOMS INDUSTRY MEDIA PLATFORM
telecomreview.com

Abu Dhabi researchers create first-of-its-kind telecom LLM

Khalifa University's 6G Centre and TII collaborate on building TelecomGPT



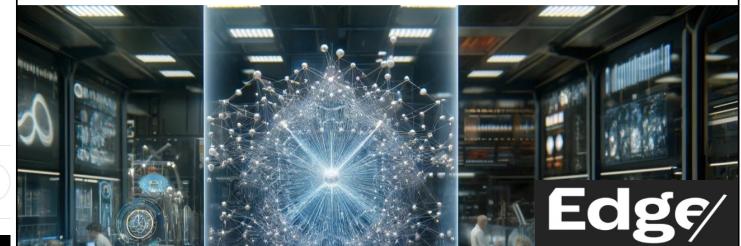
CARRINGTON MALIN
JUL 16, 2024



EMERGENT TECH July 17, 2024 09:26 AM GST

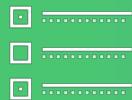
Why UAE's TelecomGPT could be a global game-changer in not just the telecom world but even the AI space

The Global Telco AI Alliance's introduction of TelecomGPT marks a pivotal moment for the ~~telecom industry~~ This advanced AI model outshines general-purpose models like GPT-4 and is AE's 6G aspirations.



Edge/

TelecomGPT OUTPERFORMS GENERAL MODELS



Multiple-Choice Telecom Questions

LLaMa 3 – 8B: 88.49%

TelecomGPT (Based on LLaMa3-8B): **94%**

Mistral – 7B: 73.25%

TelecomGPT (Based on Mistral-7B): **81.5%**



Telecom-Specific Classification

LLaMa 3 – 8B: 33.35%

TelecomGPT (Based on LLaMa3-8B): **74.3%**

Mistral – 7B: 27.84%

TelecomGPT (Based on Mistral-7B): **70.83%**



Telecom Math Modeling

LLaMa 3 – 8B: 34.45%

TelecomGPT (Based on LLaMa3-8B): **50.73%**

Mistral – 7B: 29.43%

TelecomGPT (Based on Mistral-7B): **49.26%**

Also outperforms GPT4o!

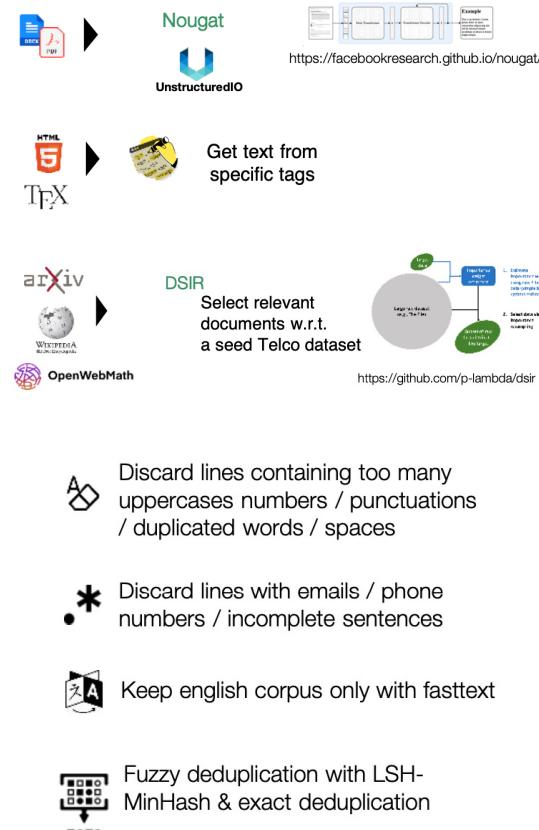
In almost all Telecom tasks although being much smaller in size!

OpenTelecom Data Pipeline

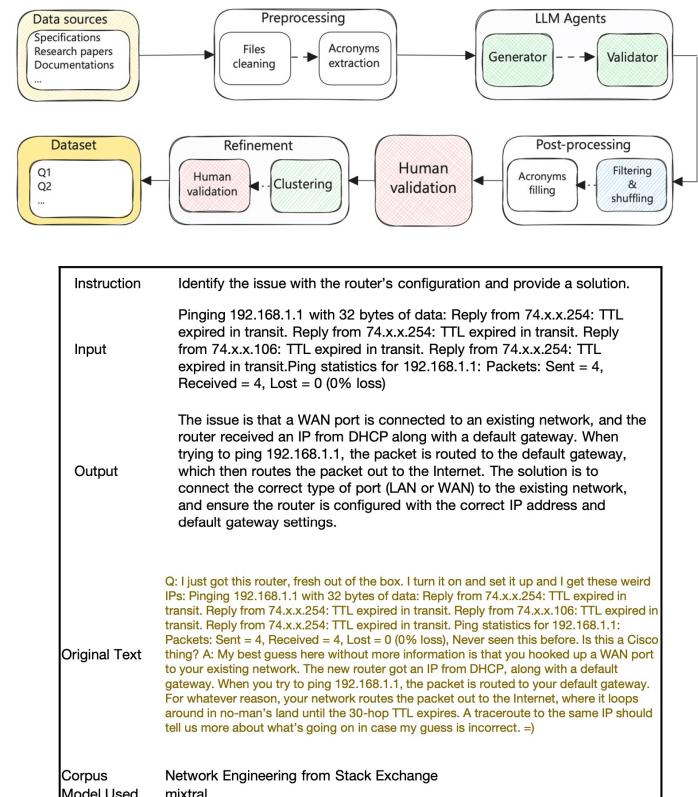
Source Collection



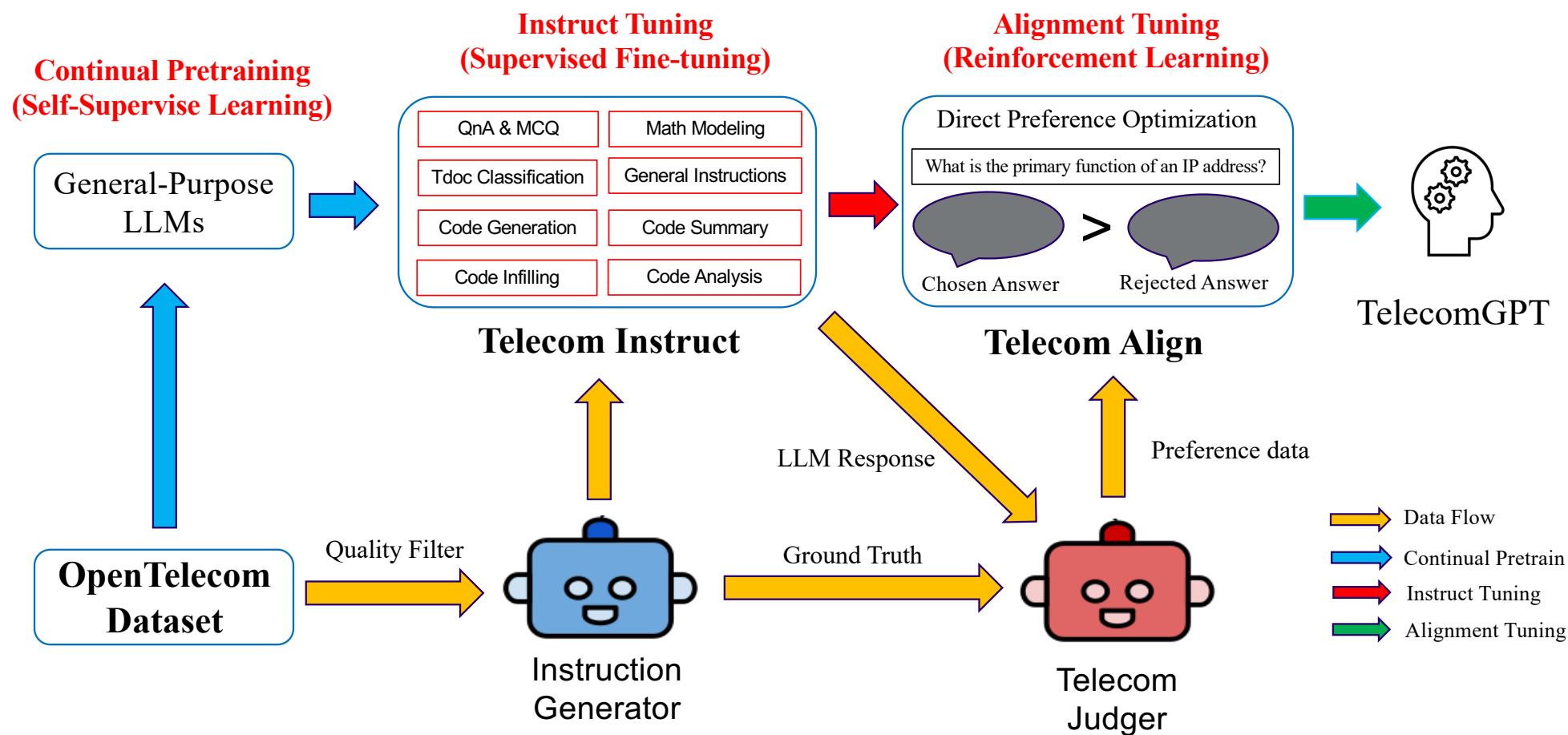
Extraction & Preprocessing



Task Specific Data Creation



TelecomGPT: Training Pipeline



OpenTelecom Dataset

Overview of pretraining Telecom dataset (M Tokens)

	Training	Validation	Percentage
3GPP standard	193	1.9	11.49
IEEE standard	7.5	0.07	0.45
Paper (arxiv)	893	9	53.17
Book	1.9	0.02	0.11
Patent (c4)	253.2	2.6	15.08
StackExchagne	51.9	0.5	3.09
Wikipedia	18.9	0.2	1.13
Code (github)	260.1	2.6	15.49
Total	1679.5	16.89	-

Training Parameters

	Contine Pretrain	Instruct Tuning	Alignment Tuning
Training	Self-Supervised Learning	Supervised Fine-Tuning	Reinfocement Learning
Epoches	1	3	1
Steps	93000	1020	180
PEFT	Full Parameters	QLoRA (2%)	QLoRA (2%)
Learning rate	2e-5 Constant	2e-5 Constant	5e-6 Cosine Decay
Training data	3263 M tokens	5M tokens	8K tokens
GPU hours (A100)	1536	46	16

Use Case & Benchmark: Telecom Knowledge Query

Queries of Telecom standards, research papers, patents, lexicons

Question	What does IM stand for, and why is it considered desirable for constellations in communication systems?
Context (optional)	<p>Option 1: Intermittent Messaging; because it allows for messages to be sent intermittently.</p> <p>Option 2: Index Modulation; because its combinatorial structure enables higher SNR efficiency, low-complexity implementation, and compatibility with existing coding and modulation techniques.</p> <p>Option 3: Instant Messaging; because it allows for real-time communication</p> <p>Option 4: Internal Modulation; because it modulates signals within the device itself.</p>
Answer	Index Modulation (IM) is desirable for communication system constellations due to its higher signal-to-noise ratio efficiency compared to conventional techniques, low-complexity implementation, and compatibility with existing coding and modulation techniques.
Category	RAN1

Potential application
(Source: AWS)

Enterprise Knowledge Search

- Challenge:** Institutional knowledge & information is often spread out across multiple sources (e.g. intranets, documents repositories, wiki sites, databases, etc.) and employees have to spend too much time searching through different systems to obtain relevant information.
- Solution:** Interactive GenAI assistant to answer employee questions, searching multiple sources of information and generating personalized answers. Solution can obtain information from knowledge bases specific to business units and access databases & APIs to generate user specific responses.
- Benefits:**
 - Reduce time to search for information and resolve issues
 - Obtain comprehensive and personalized information

Use Case & Benchmark : Telecom Task Completion and Planning

Explain techniques, identify problems, propose solutions in Telecom

Prompt	Describe the process of providing authorization to a mobile device for a pending transaction request.
Response	To provide authorization to a mobile device for a pending transaction request, the system searches the database for the mobile device's phone number, generates and sends an authorization request message to the mobile device, and awaits a reply message indicating approval or denial of the transaction from the user.

Plan tasks in Telecom specificaiton to complete a high-level instruction

Prompt	Initiate event based charging with decentralized and centralized unit determination, centralized rating
Response	<ol style="list-style-type: none">1. Request for resource usage: A request for session establishment is received in the NF (CTF). The service is configured to be authorized by the CHF to start.2. Units Determination: the NF (CTF) determines the number of units depending on the service requested by the UE in "Decentralized Units determination" scenario.3. Charging Data Request [Event, Units]: The NF (CTF) sends the request to the CHF for the service to be granted authorization, and to allow the number of units, if determined in item 2, to be rated and accounted.

Potential application
(Source: AWS)

Field Technician Assistant

- Challenge:** Field technicians encounter multiple challenges during their work in the field, from technical problems to unforeseen situations that require consulting documentation and procedures. Failure to obtain timely information in the field generate delays and repeat visits that impact costs and customer experience
- Solution:** Interactive GenAI assistant that can provide timely information to field technicians on their mobile phones, searching through technical manuals, standard procedures, etc. The GenAI assistant also has access to specific information related to the current field assignment and past field work performed to provide relevant suggestions
- Benefits:**
 - Reduce operational costs, service time and repeat visits
 - Improve field technician experience and increase success rates of field visits

Use Case & Benchmark : Telecom Code Understanding

Analyze code functionality and operational logic of Telecom features

Prompt Analyze this script's contribution to telecom workflows, specifying its primary functionalities.

```
def bsc(input_bits, p_t):
    """
    Binary Symmetric Channel.

    Parameters
    -----
    input_bits : 1D ndarray containing {0, 1}
        Input array of bits to the channel.

    p_t : float in [0, 1]
        Transition/Error probability of the channel.

    Returns
    -----
    output_bits : 1D ndarray containing {0, 1}
        Output bits from the channel.
    """
    output_bits = input_bits.copy()
    flip_locs = (random(len(output_bits)) <= p_t)
    output_bits[flip_locs] = 1 ^ output_bits[flip_locs]
    return output_bits
```

Response The script defines a function `bsc` for Binary Symmetric Channel within the telecom domain. It takes `input_bits` array of bits and `p_t` transition/error probability as parameters and returns `output_bits` after applying the channel operations. The function simulates a binary symmetric channel by flipping bits based on the given probability, which is a crucial concept in telecom communication systems for modeling channel errors and noise.

Potential application
(Source: AWS)

Developer Productivity

✓ **Challenge:** Solution developers have to spend significant amounts of time repetitive tasks like developing unit tests, creating documentation and maintaining legacy code. These tasks keep developers from focusing on building new solutions that bring business value

✓ **Solution:** Leverage the power of GenAI to generate code from plain instructions, explain, document or debug existing code and improve the quality and test coverage of solutions.

✓ **Benefits:**

- Improve developer productivity, reducing the time to create and maintain solutions
- Improve quality and consistency of code

Use Case & Benchmark : Telecom Code Generation

Generate script or function for a given task or functionality in Telecom

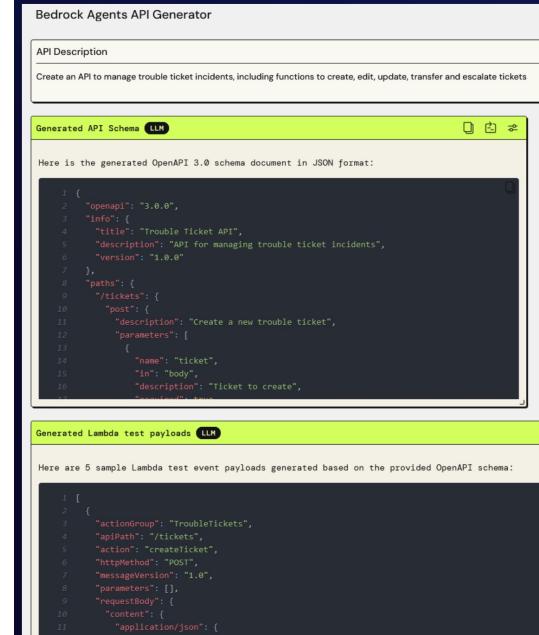
Prompt

Develop a C function that retrieves and returns the transmit (TX) power value for a wireless device. The function should update a data structure representing wireless request (wrqu) with the TX power information including the power value, the fixed state, flags, and disabled status based on the device's status. This function is intended to be used in the context of configuring and managing wireless interfaces within a telecommunications network.

Response

```
union iwreq_data *wrqu, char *extra)
{
    struct ipw_priv *priv = ieee80211_priv(dev);
    mutex_lock(&priv->mutex);
    wrqu->power.value = priv->tx_power;
    wrqu->power.fixed = 1;
    wrqu->power.flags = IW_TXPOW_DBM;
    wrqu->power.disabled = (priv->status & STATUS_RF_KILL_MASK) ? 1 : 0;
    mutex_unlock(&priv->mutex);
    IPW_DEBUG_WX("GET TX Power -> %s %d \n",
                wrqu->power.disabled ? "OFF" : "ON", wrqu->power.value);
    return 0;
}
```

Potential application (Source: AWS)



The screenshot shows the Bedrock Agents API Generator interface. It includes sections for 'API Description' (describing a trouble ticket API), 'Generated API Schema' (showing an OpenAPI 3.0 schema document in JSON format), and 'Generated Lambda test payloads' (showing 5 sample Lambda test event payloads generated based on the provided OpenAPI schema).

Generated API Schema

```
1  {
2      "openapi": "3.0.0",
3      "info": {
4          "title": "Trouble Ticket API",
5          "description": "API for managing trouble ticket incidents",
6          "version": "1.0.0"
7      },
8      "paths": {
9          "/tickets": {
10              "post": {
11                  "description": "Create a new trouble ticket",
12                  "parameters": [
13                      {
14                          "name": "ticket",
15                          "in": "body",
16                          "description": "Ticket to create"
17                      }
18                  ]
19              }
20          }
21      }
22 }
```

Generated Lambda test payloads

```
1  [
2      {
3          "actionGroup": "TroubleTickets",
4          "apiPath": "/tickets",
5          "action": "createTicket",
6          "httpMethod": "POST",
7          "messageVersion": "1.0",
8          "parameters": [],
9          "requestBody": {
10              "content": {
11                  "application/json": {
12                      "ticket": {
13                          "id": "12345678901234567890123456789012"
14                      }
15                  }
16              }
17          }
18      }
19  ]
```

Use Case & Benchmark : Telecom Math Modeling

Generate math solution according to Telecom scenario descriptions

Prompt

Please fill in the missing mathematical equation represented by < MASK > placeholder in the provided sections of a telecom-relevant research paper that discuss the system model or problem formulation.

System Model: We consider a discrete-time zero-mean jointly proper Gaussian flat-fading channel with the following input-output relation

< MASK >

with the diagonal matrices $\mathbf{H} = \text{diag}(\mathbf{h})$ and $\mathbf{X} = \text{diag}(\mathbf{x})$. Here the $\text{diag}(\cdot)$ operator generates a diagonal matrix whose diagonal elements are given by the argument vector. The vector $\mathbf{y} = [y_1, \dots, y_N]^T$ contains the channel output symbols in temporal order. Analogous, $\mathbf{x} = [x_1, \dots, x_N]^T$, $\mathbf{n} = [n_1, \dots, n_N]^T$, and $\mathbf{h} = [h_1, \dots, h_N]^T$ contain the channel input symbols, the additive noise samples and the channel fading weights. All vectors are of length N . The samples of the additive noise process are assumed to be i.i.d. zero-mean jointly proper Gaussian with variance σ_n^2 and, thus, $\mathbf{R}_n = \mathbb{E} [\mathbf{n}\mathbf{n}^H] = \sigma_n^2 \mathbf{I}_N$, with \mathbf{I}_N being the identity matrix of size $N \times N$. The channel fading process is zero-mean jointly proper Gaussian with the temporal correlation characterized by

< MASK >

Its variance is given by $r_h(0) = \sigma_h^2$. For mathematical reasons we assume that the autocorrelation function $r_h(l)$ is absolutely summable, i.e.,

< MASK >

Response

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n} = \mathbf{X}\mathbf{h} + \mathbf{n}$$

$$r_h(l) = \mathbb{E}[h_{k+l} \cdot h_k^*].$$

$$\sum_{l=-\infty}^{\infty} |r_h(l)| < \infty.$$

Router Diagnosis

Upload a photo of your router to receive a professional diagnosis and troubleshooting steps displayed in a structured, easy-to-read format.

Upload an image of the faulty router

Clear

Submit

Issue Description The image shows a Frontier router with a series of lights indicating a potential issue. The lights are red and green, which typically indicate a problem with the router's connection or configuration.

Root Cause Analysis

- o **LED Analysis**
 - o Color: Red and Green
 - o Pattern: Alternating red and green lights
 - o Indicates: Potential connection or configuration issue
- o **Error Code:** None visible
- o **Possible Cause:** The router may be experiencing a connection issue or configuration problem.

Step-by-Step Troubleshooting

1. Restart the Router

- o Action: Press the power button to turn off the router, wait for 30 seconds, and then press it again to turn it back on.
- o Details: This step is often the simplest and most effective way to resolve connectivity issues.
- o Expected Outcome: The router should boot up normally and the lights should stabilize.

2. Check Ethernet Cable

- o Action: Inspect the Ethernet cable connecting the router to the modem for any signs of damage or loose connections.
- o Details: Ensure the cable is securely plugged into both the router and modem.
- o Expected Outcome: A secure connection should be established, and the lights on the router should stabilize.

3. Check Wi-Fi Settings

- o Action: Log in to the router's web interface and check the Wi-Fi settings to ensure they are correct and not set to "off" or "disabled". +

Large Generative AI Models in Telecom (GenAINet)

Emerging Technology Initiative

The aim of this Emerging Technology Initiative (ETI) is to create a dynamic platform of research and innovation for academics, researchers, and industry leaders to advance the research on large generative AI in Telecom, through collaborative efforts across various disciplines, including mathematics, information theory, wireless communications, signal processing, networking, artificial intelligence, and more. Generative AI, a cutting-edge technology within AI, is anticipated to reshape the future of telecommunications by enhancing network capabilities, optimizing operations, and creating new opportunities for seamless connectivity. Alternatively, future network generations are envisioned to play an essential role in enabling collective intelligence through realizing multi-agent generative AI.

Accordingly, this ETI is committed to exploring the transformative potential of Generative AI in Telecom through, 1) supporting the development of novel architectures tailored for Telecom within large Generative AI models, 2) curating dataset libraries for training and fine-tuning Telecom-specific GenAI models, 3) developing and implementing sophisticated fine-tuning strategies, 4) enhancing communication schemes and protocols for generative AI agents, 5) the development of theoretical frameworks for explainable GenAI in Telecom, and 6) establishing benchmarking schemes and performance metrics to rigorously evaluate the effectiveness of large Telecom GenAI models.



General Chair | Merouane Debbah

Academic Chair | Tingting Yang

Industry Chair | Lina Bariah

Vice Chairs | Mehdi Bennis, Deniz Gunduz, Kenza Hamidouche, Jean-Claude Belfiore

Workshops, Special Sessions, and Conferences | Samson Lasaulce, David Lopez-Perez

Special Issues in Top-Tier IEEE Journals | Abdelmalik Bachir, Carlo Fischione, Emilio Calvanese Strinati

Tutorials, Invited Talks, and Industry Panels | Christina Chaccour, Xingqin Lin, Juan Deng

Dataset and Competitions | Qiyang Zhao, Igor Carron, Antonio De Domenico

Industry Activities and Standardization Officer | Chenghui Peng, Fathi Abdeldayem, Markus Mueck

Secretary | Li Sun

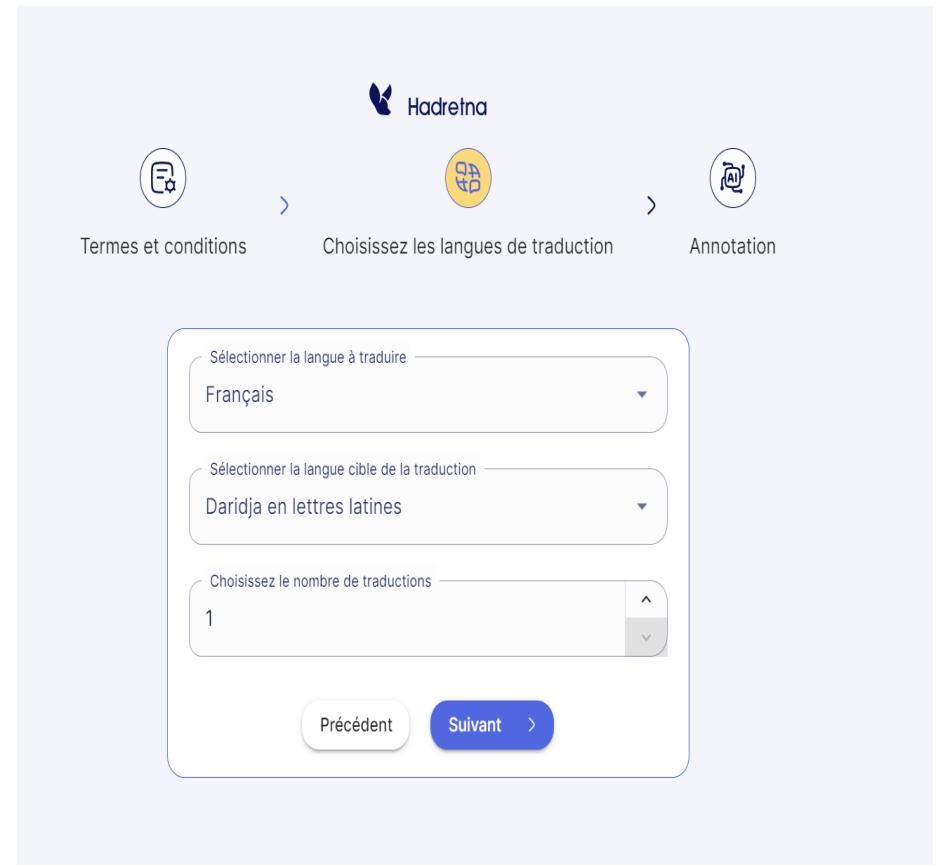
Hadratena



Hadretna

Hadretna houwa projet ta3 Ai Bach yssahel l'utilisation ta3 lahajate ta3 l'Afrique du nord : Daridja. Le but houwa nssahlou lel les personnes ta3 l'Afrique du Nord Bach yalha9elhom Kamel les information ta3 le monde bel el-lahdja li yahder biha.

Discutez avec Hadretna >



Hadretna

Terminé

Choisissez les langues de traduction

Annotation

Sélectionner la langue à traduire
Français

Sélectionner la langue cible de la traduction
Daridja en lettres latines

Choisissez le nombre de traductions
1

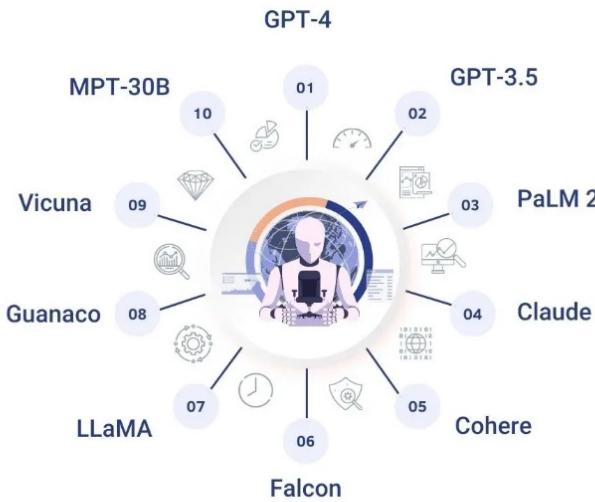
Précédent > Suivant >

Talk Outline

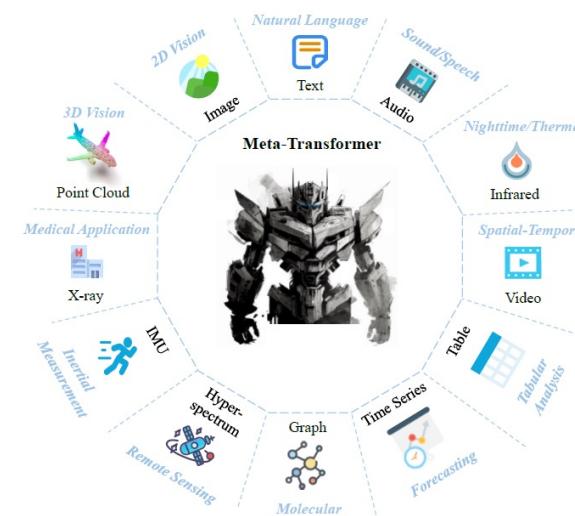
- A General Introduction
- Our Work on Generative AI
- Thoughts on next steps in Generative AI

Where are we now?

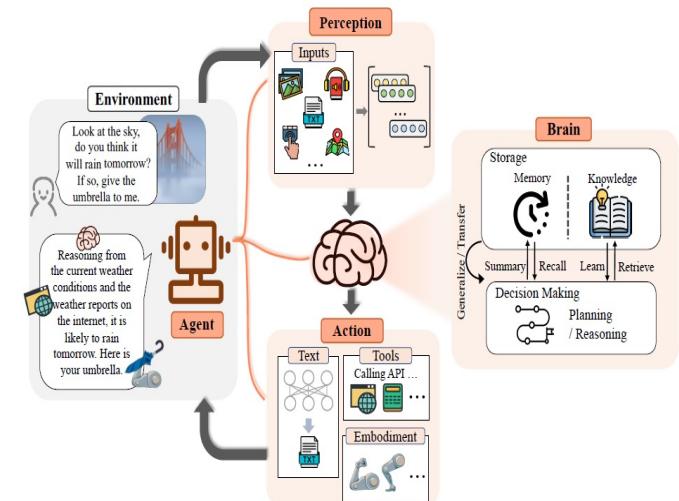
BEYOND GPT



BIG DATA

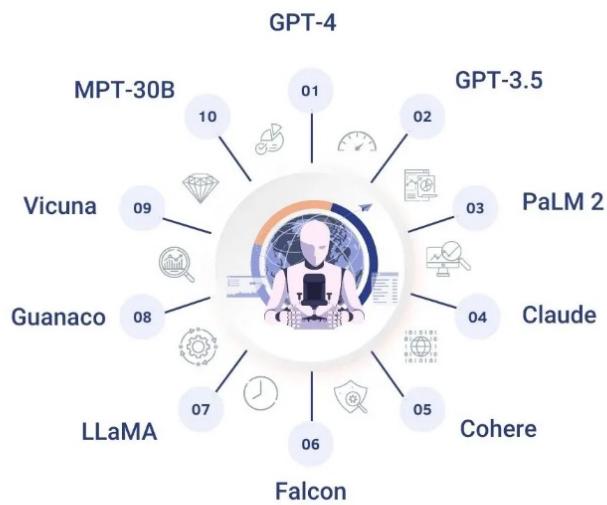


AUTONOMY



Where are we now?

BEYOND GPT



Beyond GPT

	Model size	Data size	Mixture-of-experts	State space models
Training	Easier to distribute trivial to do	Longer training hard to source	~Equivalent to larger model	~Equivalent?
Inference	Expensive	Same cost!	Only pay for active params	Better scalability?

Beyond GPT

Were RNNs All We Needed?

Leo Feng
Mila – Université de Montréal & Borealis AI
leo.feng@mila.quebec

Frederick Tung
Borealis AI
frederick.tung@borealisai.com

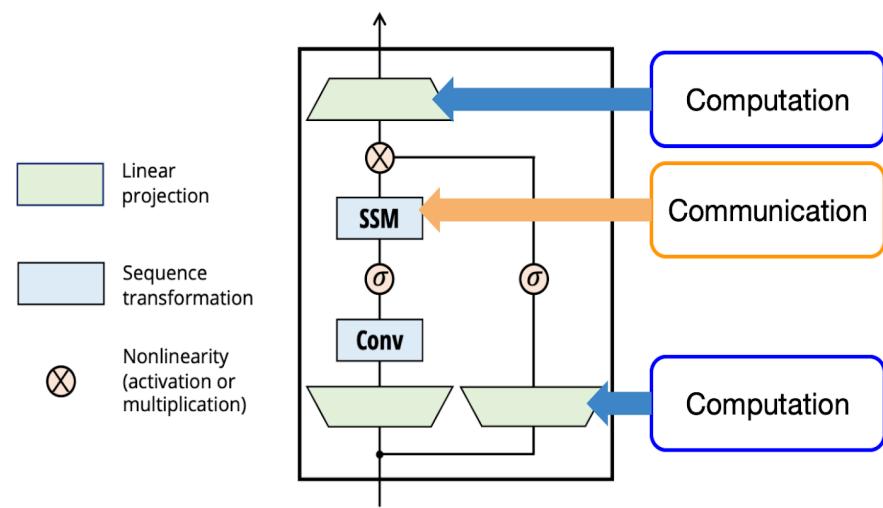
Mohamed Osama Ahmed
Borealis AI
mohamed.o.ahmed@borealisai.com

Yoshua Bengio
Mila – Université de Montréal
yoshua.bengio@mila.quebec

Hossein Hajimirsadeghi
Borealis AI
hossein.hajimirsadeghi@borealisai.com

Abstract

The scalability limitations of Transformers regarding sequence length have renewed interest in recurrent sequence models that are parallelizable during training. As a result, many novel recurrent architectures, such as S4, Mamba, and Aaren, have been proposed that achieve comparable performance. In this work, we revisit traditional recurrent neural networks (RNNs) from over a decade ago: LSTMs (1997) and GRUs (2014). While these models were slow due to requiring to backpropagate through time (BPTT), we show that by removing their hidden state dependencies from their input, forget, and update gates, LSTMs and GRUs no longer need to BPTT and can be efficiently trained in parallel. Building on this, we introduce minimal versions (minLSTMs and minGRUs) that (1) use significantly fewer parameters than their traditional counterparts and (2) are fully parallelizable during training ($175\times$ faster for a sequence of length 512). Lastly, we show that these stripped-down versions of decade-old RNNs match the empirical performance of recent sequence models.



Mamba

Pre-train less and Infer more



@DrJimFan

Beyond GPT

DIFFERENTIAL TRANSFORMER

Tianzhu Ye^{*†‡} Li Dong^{*†} Yuqing Xia^{*†} Yutao Sun^{*†‡}

Yi Zhu[†] Gao Huang[†] Furu Wei^{†○}

[†] Microsoft Research [†] Tsinghua University

<https://aka.ms/GeneralAI>

Abstract

Transformer tends to overallocate attention to irrelevant context. In this work, we introduce DIFF Transformer, which amplifies attention to the relevant context while canceling noise. Specifically, the differential attention mechanism calculates attention scores as the difference between two separate softmax attention maps. The subtraction cancels noise, promoting the emergence of sparse attention patterns. Experimental results on language modeling show that DIFF Transformer outperforms Transformer in various settings of scaling up model size and training tokens. More intriguingly, it offers notable advantages in practical applications, such as long-context modeling, key information retrieval, hallucination mitigation, in-context learning, and reduction of activation outliers. By being less distracted by irrelevant context, DIFF Transformer can mitigate hallucination in question answering and text summarization. For in-context learning, DIFF Transformer not only enhances accuracy but is also more robust to order permutation, which was considered as a chronic robustness issue. The results position DIFF Transformer as a highly effective and promising architecture to advance large language models.

Will we run out of electricity?

Year	OOMs	H100s-equivalent	Cost	Power	Power reference class
2022	~GPT-4 cluster	~10k	~\$500M	~10 MW	~10,000 average homes
~2024	+1 OOM	~100k	\$billions	~100MW	~100,000 homes
~2026	+2 OOMs	~1M	\$10s of billions	~1 GW	The Hoover Dam, or a large nuclear reactor
~2028	+3 OOMs	~10M	\$100s of billions	~10 GW	A small/medium US state
~2030	+4 OOMs	~100M	\$1T+	~100GW	>20% of US electricity production

Energy Efficient Models

ADDITION IS ALL YOU NEED FOR ENERGY-EFFICIENT LANGUAGE MODELS

Hongyin Luo & Wei Sun
BitEnergy AI, Inc.
Cambridge, MA 02142, USA
`{hongyin,wei}@bitenergy.ai`

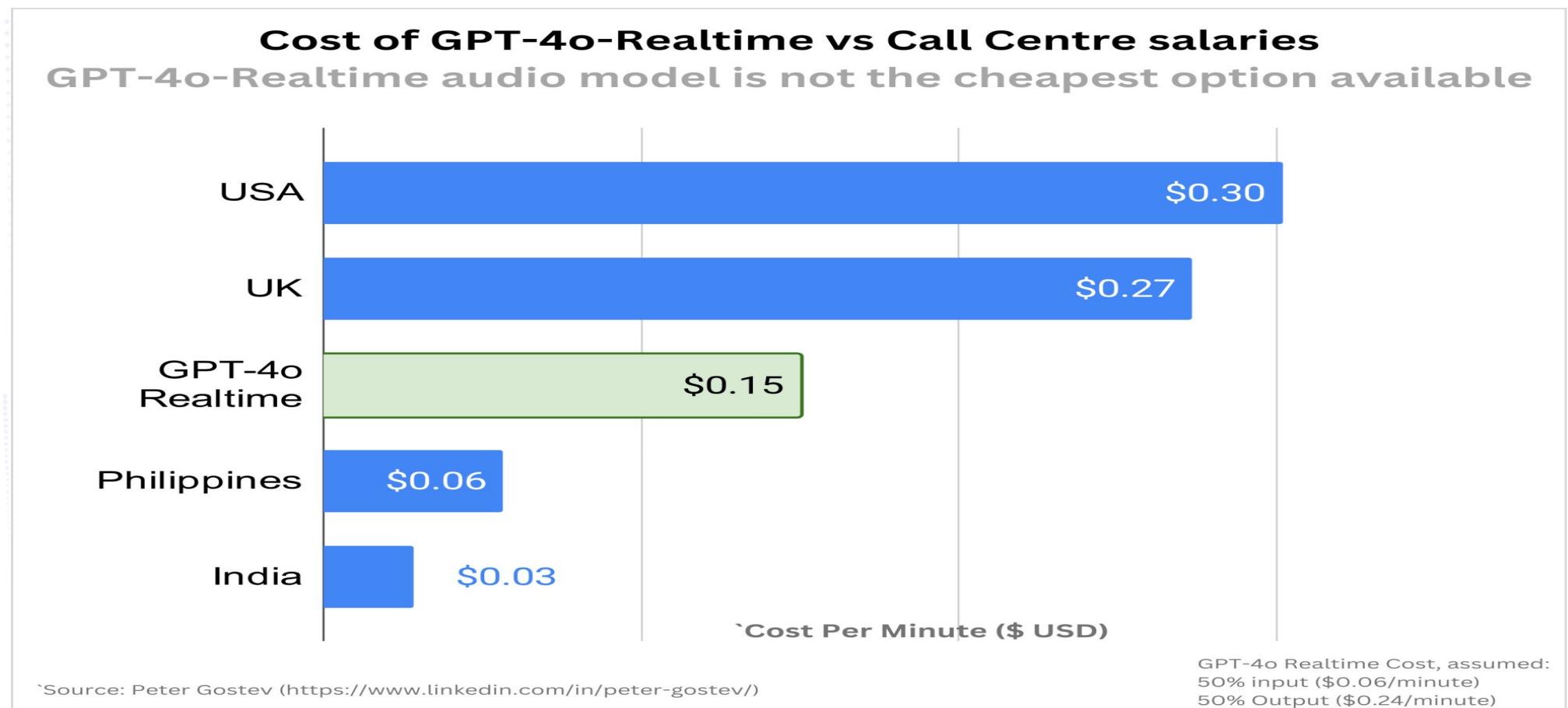
ABSTRACT

Large neural networks spend most computation on floating point tensor multiplications. In this work, we find that a floating point multiplier can be approximated by one integer adder with high precision. We propose the linear-complexity multiplication (\mathcal{L} -Mul) algorithm that approximates floating point number multiplication with integer addition operations. The new algorithm costs significantly less computation resource than 8-bit floating point multiplication but achieves higher precision. Compared to 8-bit floating point multiplications, the proposed method achieves higher precision but consumes significantly less bit-level computation. Since multiplying floating point numbers requires substantially higher energy compared to integer addition operations, applying the \mathcal{L} -Mul operation in tensor processing hardware can potentially reduce 95% energy cost by element-wise floating point tensor multiplications and 80% energy cost of dot products. We calculated the theoretical error expectation of \mathcal{L} -Mul, and evaluated the algorithm on a wide range of textual, visual, and symbolic tasks, including natural language understanding, structural reasoning, mathematics, and commonsense question answering. Our numerical analysis experiments agree with the theoretical error estimation, which indicates that \mathcal{L} -Mul with 4-bit mantissa achieves comparable precision as `float8_e4m3` multiplications, and \mathcal{L} -Mul with 3-bit mantissa outperforms `float8_e5m2`. Evaluation results on popular benchmarks show that directly applying \mathcal{L} -Mul to the attention mechanism is almost lossless. We further show that replacing all floating point multiplications with 3-bit mantissa \mathcal{L} -Mul in a transformer model achieves equivalent precision as using `float8_e4m3` as accumulation precision in both fine-tuning and inference.

1 INTRODUCTION

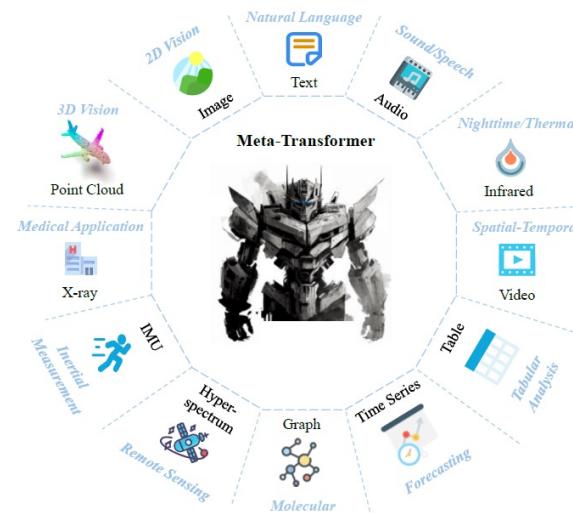
Modern artificial intelligence (AI) systems are significant energy consumers. Because of the large scale computation needed for neural network inference, AI applications based on such models are consuming a considerable amount of electricity resource. Reportly, the average electricity consumption of ChatGPT service in early 2023 was 564 MWh per day, equivalent to the total daily electricity usage of 18,000 families in the United States¹. It is estimated that Google's AI service could consume as much electricity as Ireland (29.3 TWh per year) in the worst-case scenario (de Vries, 2023).

Inference Cost?



Where are we now?

BIG DATA



Are we doing the right evaluations?

Pretraining

SEC Charges LA-Based Media and Entertainment Co. Impact Theory for Unregistered Offering of NFTs

FOR IMMEDIATE RELEASE
2023-163

Washington D.C., Aug. 28, 2023 — The Securities and Exchange Commission today charged Impact Theory, LLC, a media and entertainment company headquartered in Los Angeles, with conducting an unregistered offering of crypto asset securities in the form of purported non-fungible tokens (NFTs). Impact Theory raised approximately \$30 million from hundreds of investors, including investors across the United States, through the offering.

According to the SEC's order, from October to December 2021, Impact Theory offered and sold three tiers of NFTs, known as Founder's Keys, which Impact Theory called "Legendary," "Heroic," and "Relentless." The order finds that Impact Theory encouraged potential investors to view the purchase of a Founder's Key as an investment into the business, stating that investors would profit from their purchases if Impact Theory was successful in its efforts. Among other things, Impact Theory emphasized that it was "trying to build the next Disney," and, if successful, it would deliver "tremendous value" to Founder's Key purchasers. The order finds that the NFTs offered and sold to investors were investment contracts and therefore securities. Accordingly, Impact Theory violated the federal securities laws by offering and selling these crypto asset securities to the public in an unregistered offering that was not otherwise exempt from registration.

Predict the next word,
predominantly for web data



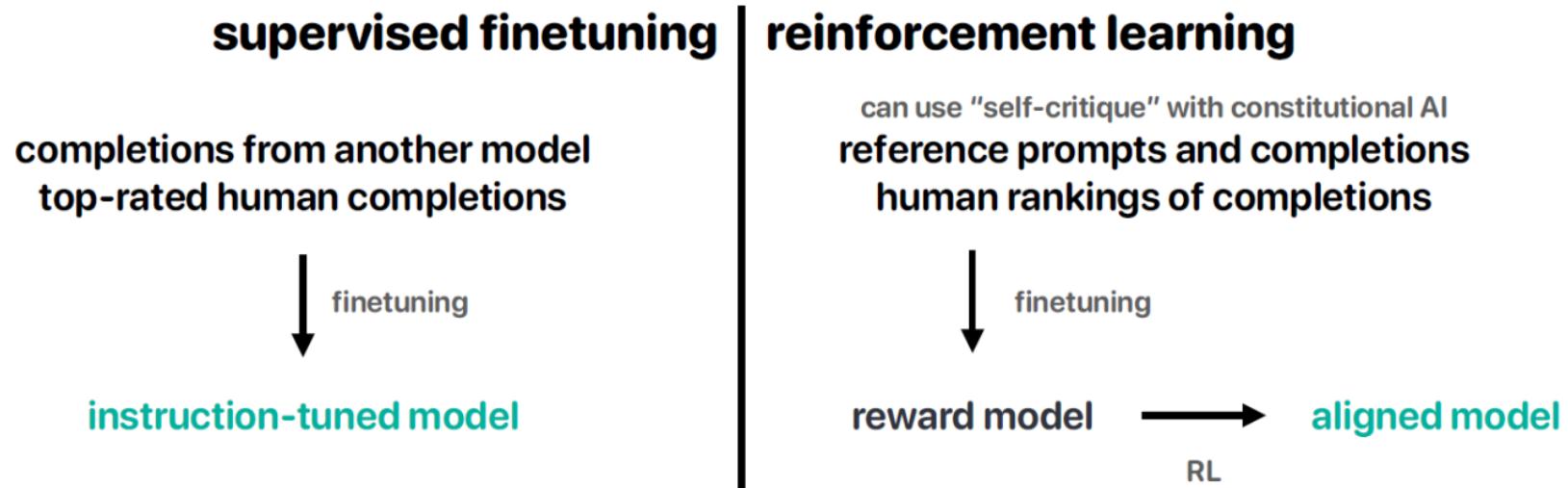
Downstream use

- Can you list 5 things to do in Reykjavik?
- Can you make my e-mail more professional?
- When should I use a dataclass in Python?

Follow users' instructions,
in a helpful, harmless, honest way.

Downstream Adaptation defines Language Models!

Knowledge is likely already there,
but need to **steer** the model to find it!



Evaluations are broken!

LLM evaluations are currently in crisis!

You **CAN'T** compare pretrained vs finetuned/RL'd models!

You (mostly) **SHOULDN'T** evaluate finetuned/RL'd models with NLP task!

(😊 good news: big aggregated benchmarks are probably totally okay as a general measure for pretraining!)

Model-based evaluations are promising, but still warrant doubt.

(issues with style, tasks beyond classic requests, and reproducibility)

Human-based evaluations are the gold standard, but expensive!

Will we run out of Data?

Will we run out of data? Limits of LLM scaling based on human-generated data

Pablo Villalobos¹ Anson Ho¹ Jaime Sevilla^{1,2} Tamay Besiroglu^{1,3} Lennart Heim^{1,4} Marius Hobbahn^{1,5}

Abstract

We investigate the potential constraints on LLM scaling posed by the availability of public human-generated text data. We forecast the growing demand for training data based on current trends and estimate the total stock of public human text data. Our findings indicate that if current LLM development trends continue, models will be trained on datasets roughly equal in size to the available stock of public human text data between 2026 and 2032, or slightly earlier if models are overtrained. We explore how progress in language modeling can continue when human-generated text datasets cannot be scaled any further. We argue that synthetic data generation, transfer learning from data-

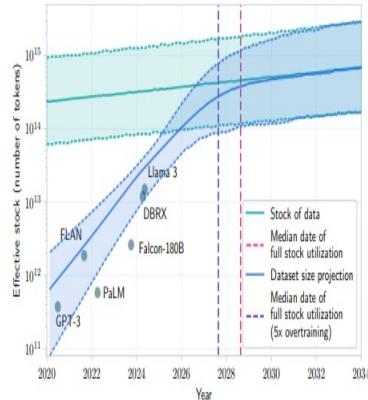
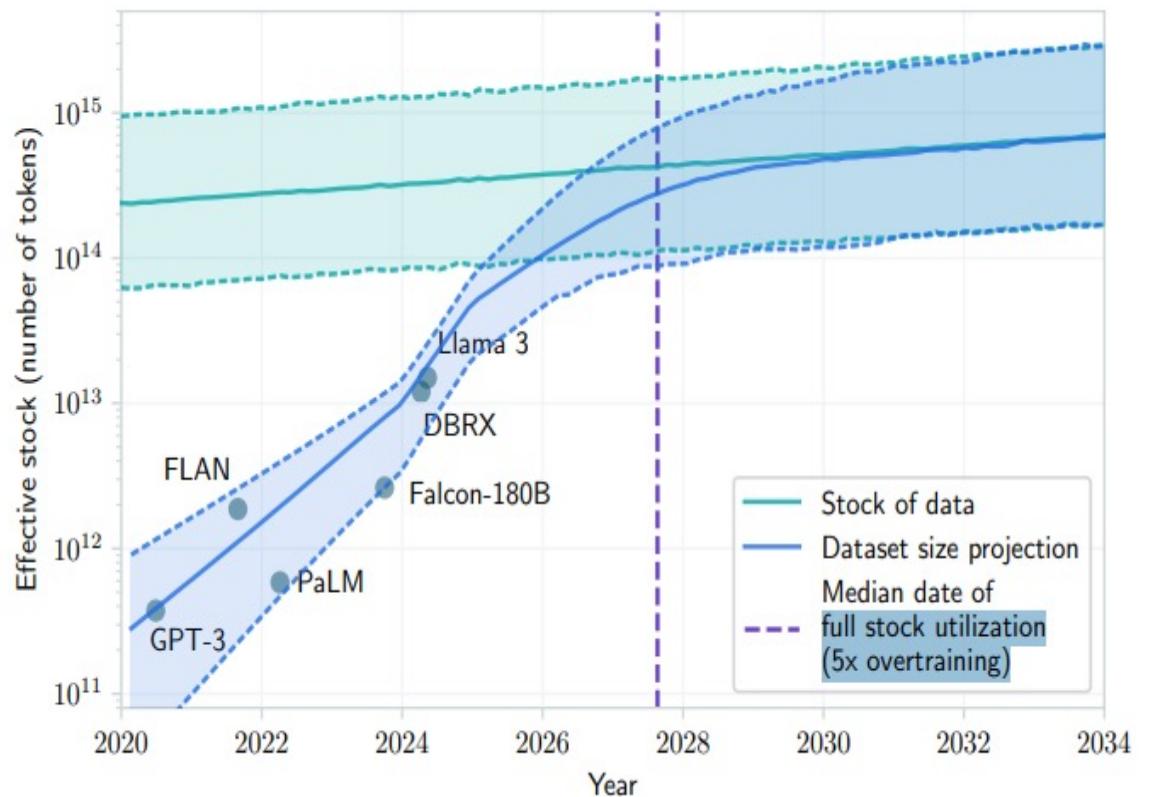


Figure 1. Projections of the effective stock of human-generated public text and dataset sizes used to train notable LLMs. The intersection of the stock and dataset size projection lines indicate



What about training on synthetic Data? Model Collapse...

How Bad is Training on Synthetic Data? A Statistical Analysis of Language Model Collapse

Mohamed El Amine Seddik
Technology Innovation Institute
Abu Dhabi, UAE
mohamed.seddik@tii.ae

Suei-Wen Chen
NYU Abu Dhabi
Abu Dhabi, UAE
swc435@nyu.edu

Soufiane Hayou
Simons Institute
Berkeley, USA
hayou@berkeley.edu

Pierre Youssef
NYU Abu Dhabi
Abu Dhabi, UAE
yp27@nyu.edu

Merouane Debbah
Khalifa University of Science and Technology
Abu Dhabi, UAE
merouane.debbah@ku.ac.ae

Abstract

The phenomenon of model collapse, introduced in (Shumailov et al., 2023), refers to the deterioration in performance that occurs when new models are trained on synthetic data generated from previously trained models. This recursive training loop makes the tails of the original distribution disappear, thereby making future-generation models forget about the initial (real) distribution. With the aim of rigorously understanding model collapse in language models, we consider in this paper a statistical model that allows us to characterize the impact of various recursive training scenarios. Specifically, we demonstrate that model collapse cannot be avoided when training solely on synthetic data. However, when mixing both real and synthetic data, we provide an estimate of a maximal amount of synthetic data below which model collapse can eventually be avoided. Our theoretical conclusions are further supported by empirical validations.

LLM for IoT

Generative AI for RF Sensing in IoT systems

Li Wang, Chao Zhang, Qiyang Zhao, Hang Zou,
Samson Lasaulce, Giuseppe Valenzise, Zhuo He, and Merouane Debbah

Abstract—The development of wireless sensing technologies, using signals such as Wi-Fi, infrared, and RF to gather environmental data, has significantly advanced within Internet of Things (IoT) systems. Among these, Radio Frequency (RF) sensing stands out for its cost-effective and non-intrusive monitoring of human activities and environmental changes. However, traditional RF sensing methods face significant challenges, including noise, interference, incomplete data, and high deployment costs, which limit their effectiveness and scalability. This paper investigates the potential of Generative AI (GenAI) to overcome these limitations within the IoT ecosystem. We provide a comprehensive review of state-of-the-art GenAI techniques, focusing on their application to RF sensing problems. By generating high-quality synthetic data, enhancing signal quality, and integrating multi-modal data, GenAI offers robust solutions for RF environment reconstruction, localization, and imaging. Additionally, GenAI's ability to generalize enables IoT devices to adapt to new environments and unseen tasks, improving their efficiency and performance. The main contributions of this article include a detailed analysis of the challenges in RF sensing, the presentation of innovative GenAI-based solutions, and the proposal of a unified framework for diverse RF sensing tasks. Through case studies, we demonstrate the effectiveness of integrating GenAI models, leading to advanced, scalable, and intelligent IoT systems.

Index Terms—Generative AI, RF sensing, cross-modal estimation, multi-modal fusion, large language models.

I. INTRODUCTION

WITH the development of the Internet of Things (IoT), many kinds of wireless sensing signals (e.g., Wi-Fi, Infrared images, visible images, Radio Frequency (RF) signal) are filling our living and working spaces nowadays. Recently, researchers have also utilized RF signals to capture events in the IoT environment (i.e., RF sensing). While RF signals are transmitted, reflected, blocked, and scattered by objects like walls, furniture, vehicles, and human bodies, it is possible to extract useful information, such as position, movement direction, speed, and vital signs of a human subject, from received RF signals. Unlike traditional hardware sensors, RF sensing provides users with low-cost and unobtrusive services. Furthermore, due to the broadcast nature of RF signals, RF sensing can be used not only to monitor multiple subjects, but also to capture changes in the environment over a large area

costs of deploying and maintaining extensive sensor networks make large-scale implementations expensive. Additionally, unstable environments cause signal weakening and multipath propagation, reducing reliability. These challenges necessitate advanced solutions like Generative AI (GenAI) to enhance the robustness, efficiency, and scalability of IoT systems.

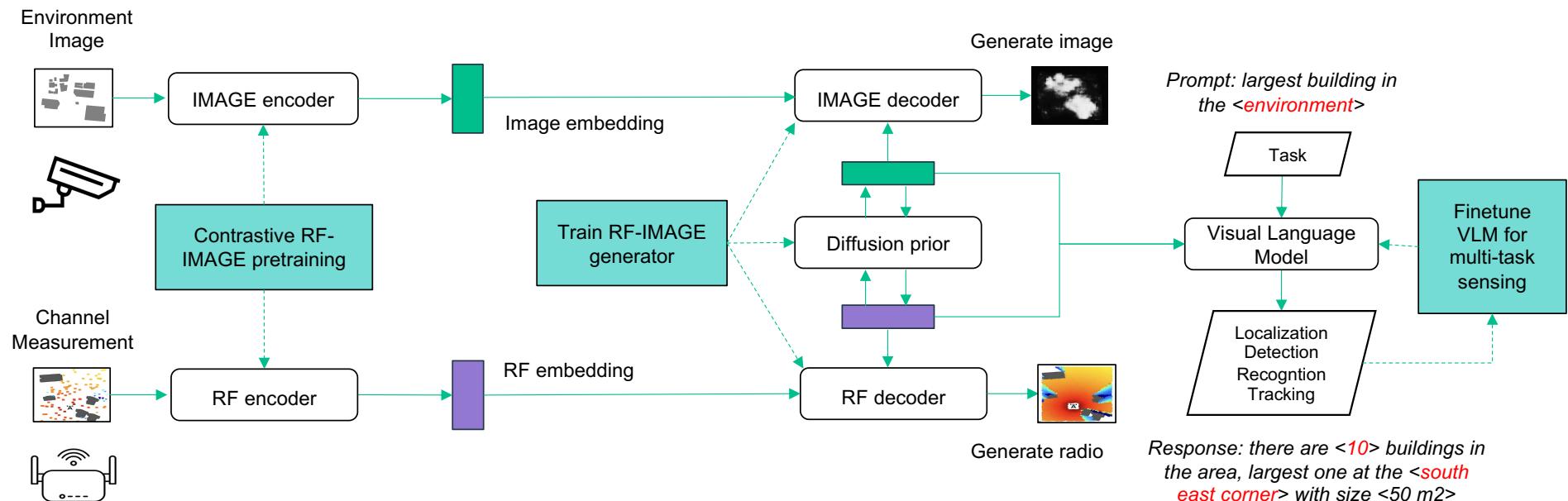
GenAI refer to neural network models designed to generate new data similar to a given dataset, including Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), Autoregressive Models, flow-Based Models, Diffusion Models (DMs), and Transformer-based Large Language Models (LLMs). These techniques offer significant advantages in data-intensive applications by creating high-quality synthetic data, improving data quality through denoising, and filling in missing values. Generative AI is particularly effective in both cross-modal and multi-modal applications: integrating diverse data types into unified representations for better decision-making and translating information between modalities to enhance robustness. This capability supports innovative IoT applications, smart cities, healthcare, and autonomous systems, showcasing generative AI's transformative potential.

GenAI's ability to enhance data quality and integrate various data types makes it ideal for IoT applications, which require universality. With the advent of smarter devices, advanced sensors, and enhanced connectivity technologies like 5G and 6G, IoT systems can greatly benefit from GenAI. It extends conventional deep learning to manage diverse and unforeseen tasks with limited data and resources. GenAI's generalization capability is crucial for IoT devices to adapt to new environments and tasks. Additionally, GenAI's natural language processing enhances multi-modal sensing by integrating text, audio, and visual data, creating more comprehensive and intelligent IoT systems.

As shown in Fig. 1, we discuss the main challenges in RF sensing applications and explore how GenAI can address these issues using unimodal and multi-modal datasets, including reviewing the most relevant works and proposing feasible solutions for the potential use of GenAI. The main contributions of this article can be summarized as follows:

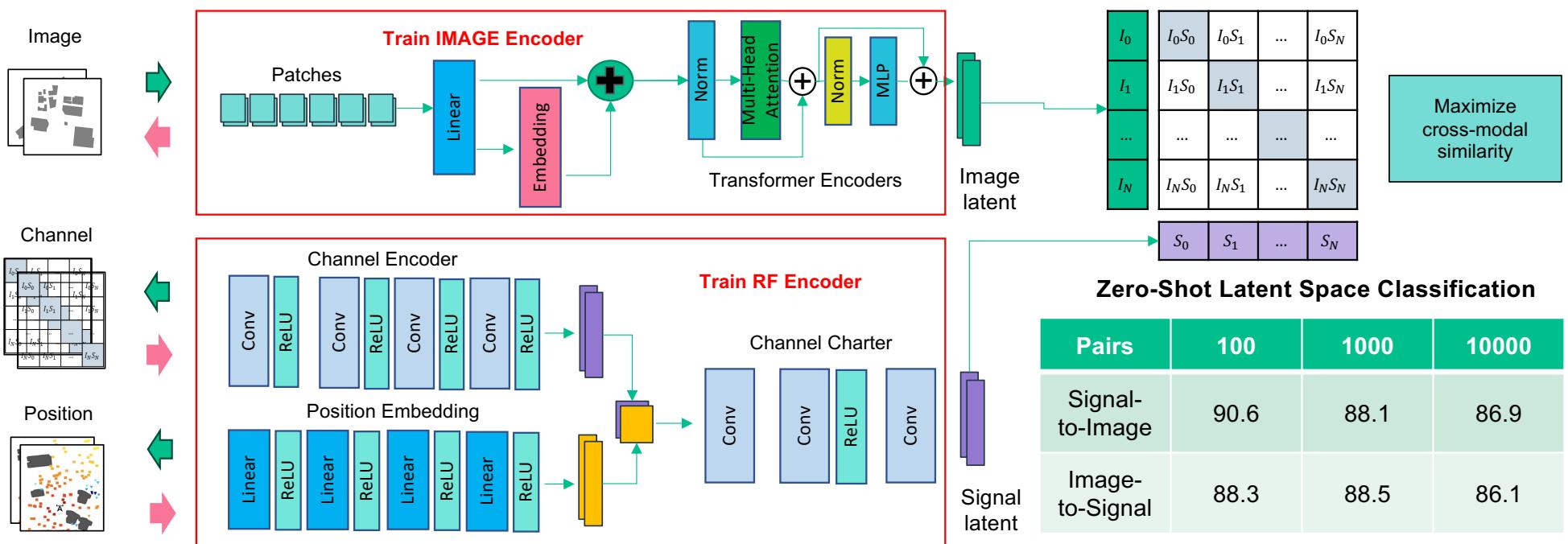
Multi-modality

- RF-Visual-Language model generalize different wireless sensing tasks with prompts
 - Contrastive cross modality pretraining to connect RF and image on a common latent space
 - Cross-attention embeds RF, image, text with LLM to generate objects, location from prompt

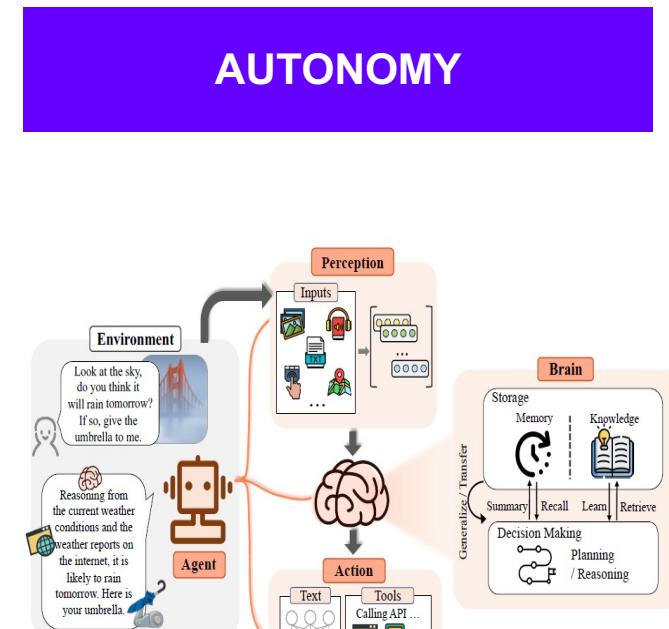


Multi-Modality

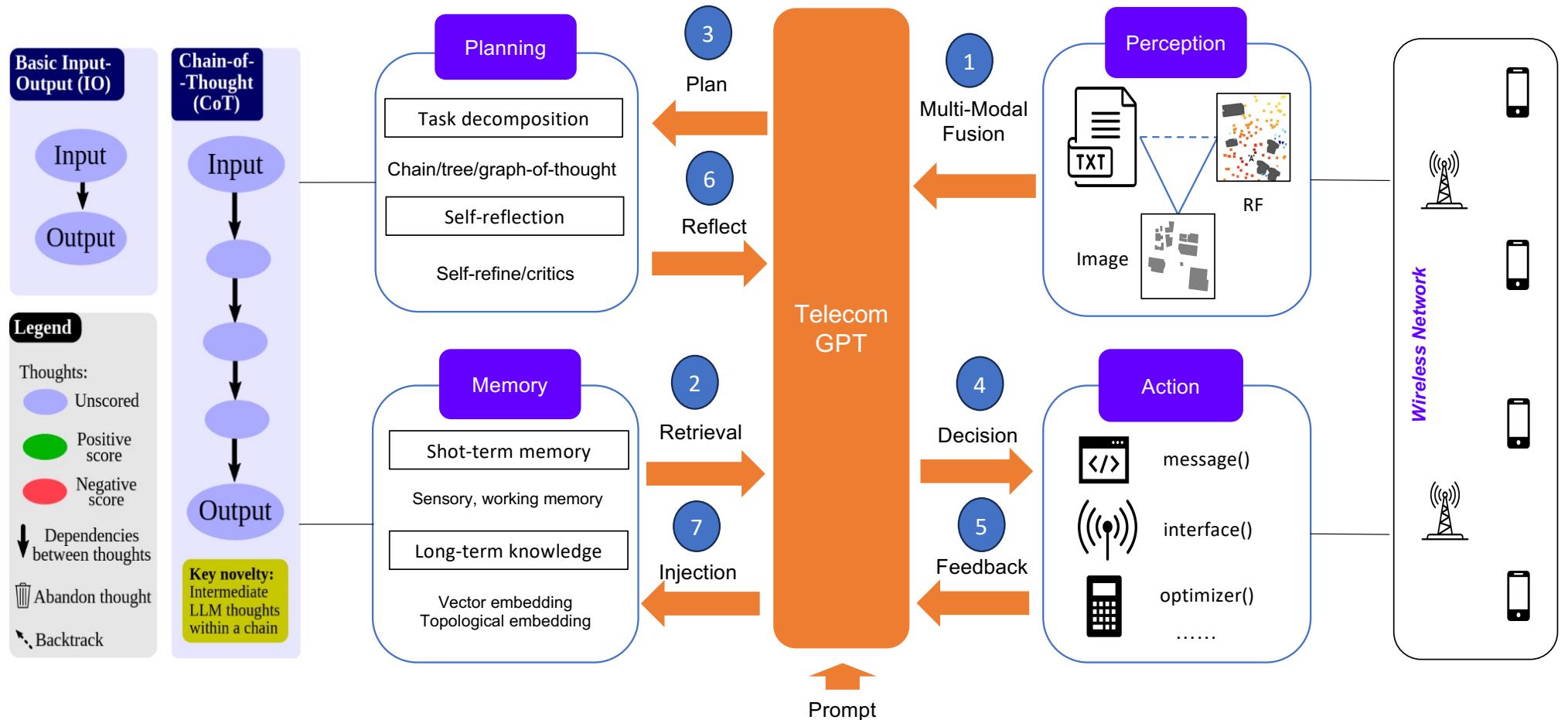
- Contrastive cross-modality pretraining
 - Maximize latent similarity of RF and image pairs sampled in the same environment
 - Capable of zero-shot classification on larger datasets without specific fine-tuning



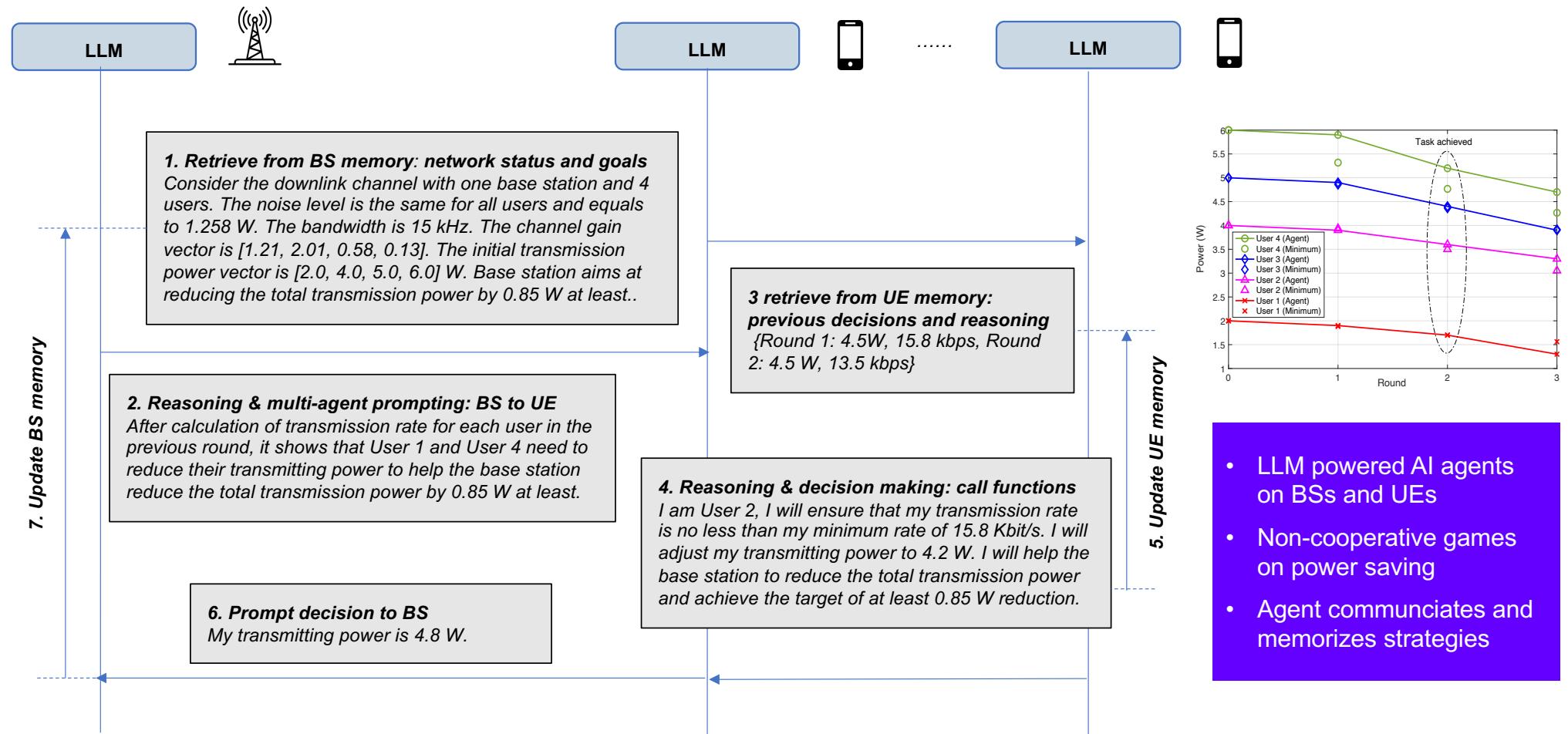
Where are we now?



TelecomGPT 3.0: Telecom AI Agent Solver

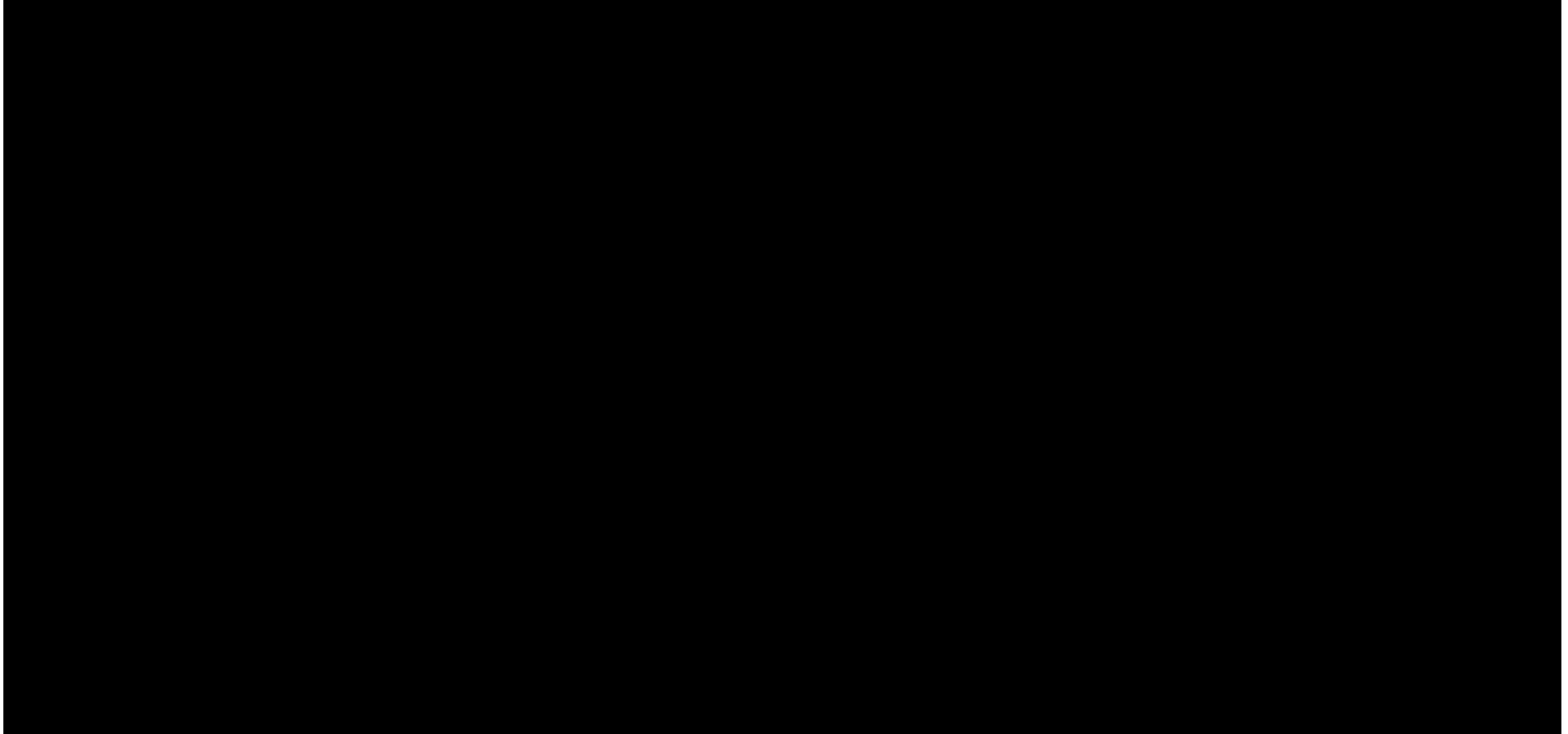


Telecom AI Agent



- LLM powered AI agents on BSs and UEs
- Non-cooperative games on power saving
- Agent communicates and memorizes strategies

Telecom AI agents: end of apps...and the rise of solvers



Thank you

