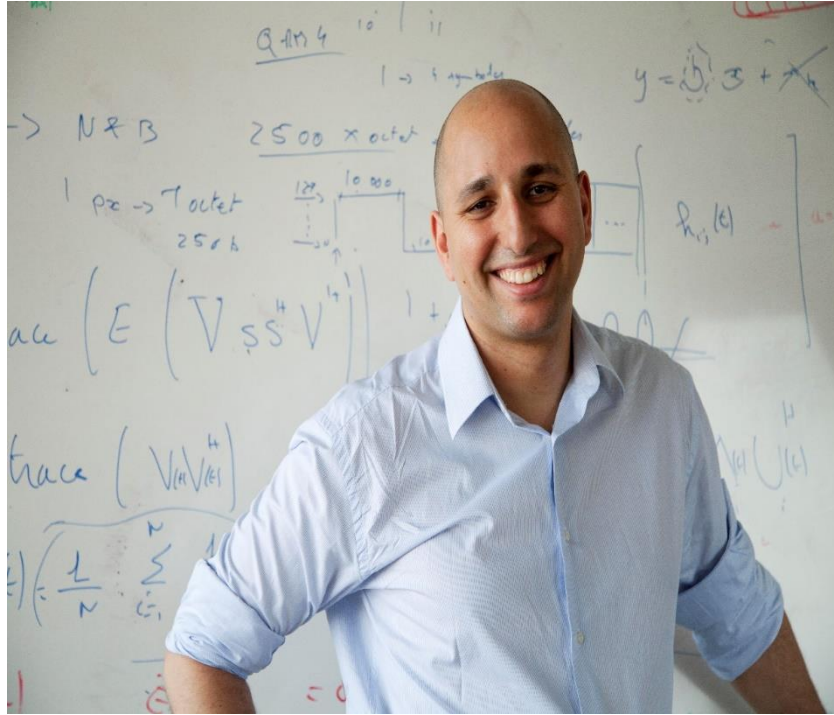


# Energy Bill of Extreme-scale language models: From theory to practice

Prof. Merouane Debbah

Joint work with I. Lakim, E. Almazrouei, I. Abu Alhaol and J. Launay



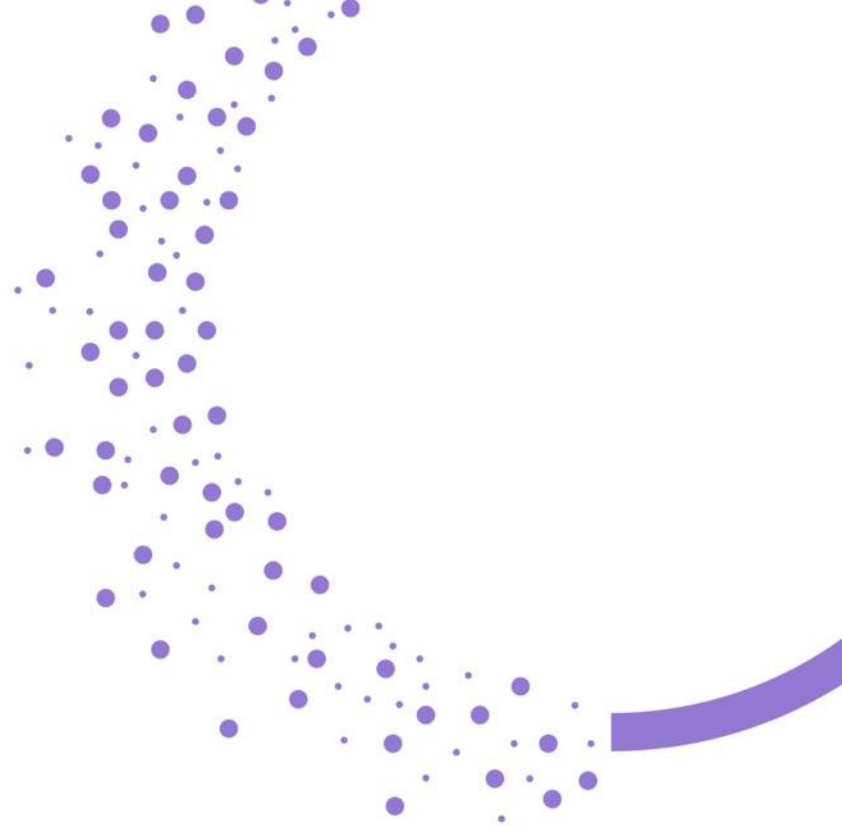
## About the Researcher

- Chief Researcher at the Technology Innovation Institute
- IEEE, EURASIP and WWRF Fellow
- Citations: 47000+, h-index: 99
- More than 20 Best papers Awards
- IEEE Signal Processing Society Distinguished Industry Speaker (2021-2022)
- Field of Research: 6G and AI

# Foundation models

Artificial intelligence  
Research Center

TII – Technology Innovation Institute

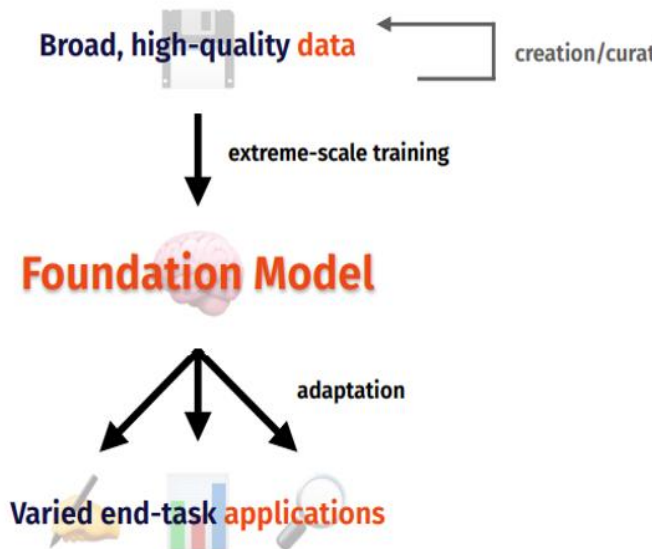


# Foundation Models

A paradigm shift in machine learning

Many Use cases beyond NLP

 Extreme-scale is changing everything, everywhere:



On the opportunities and risks of foundation models,  
Bommasani et al.



 Computer vision



 Proteins

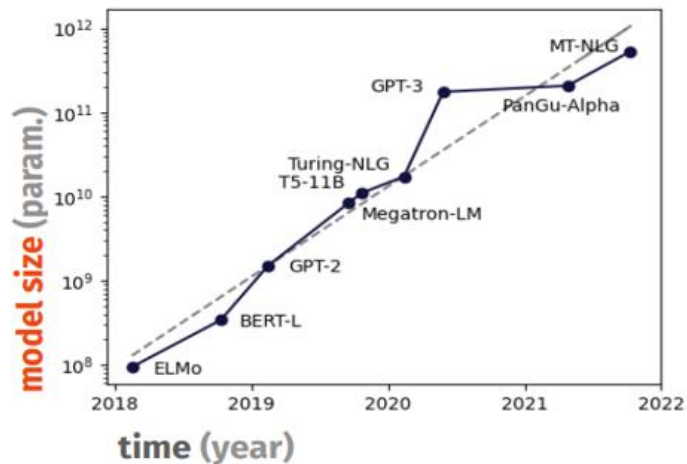


ProGen/ProtTrans (1-10B)

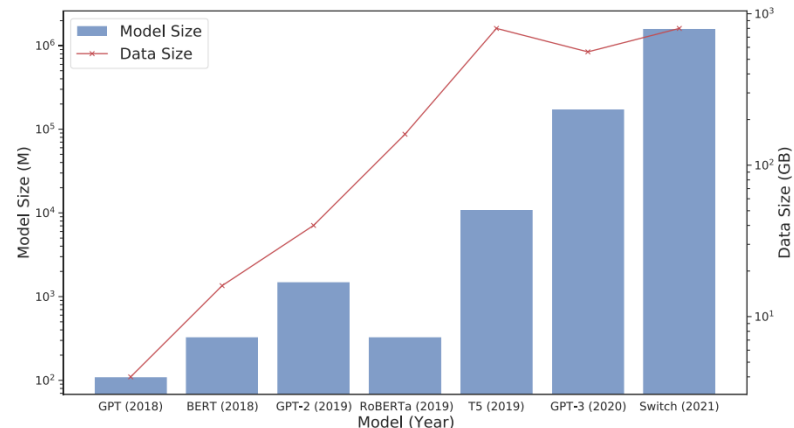


# Extreme-scale models for NLP

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

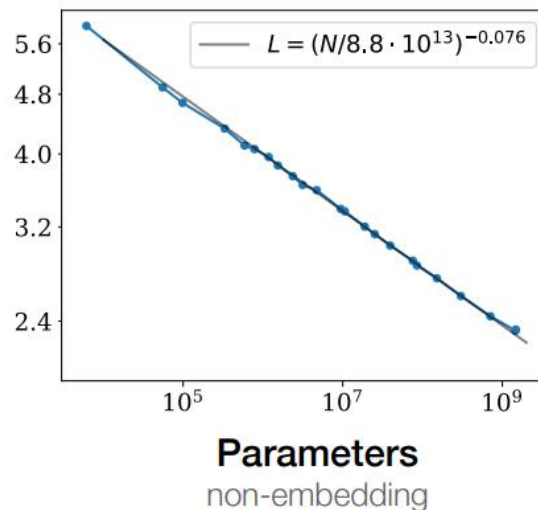
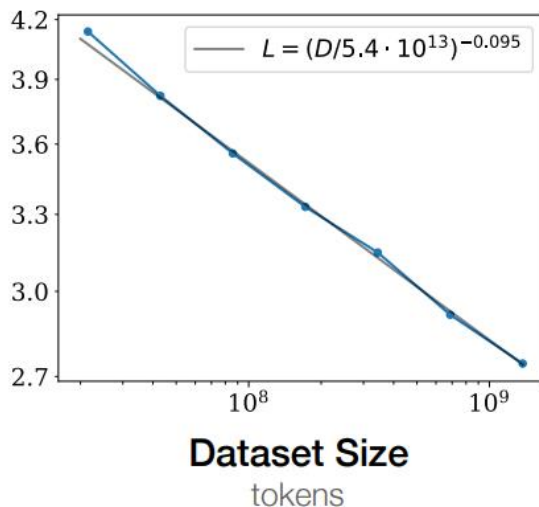
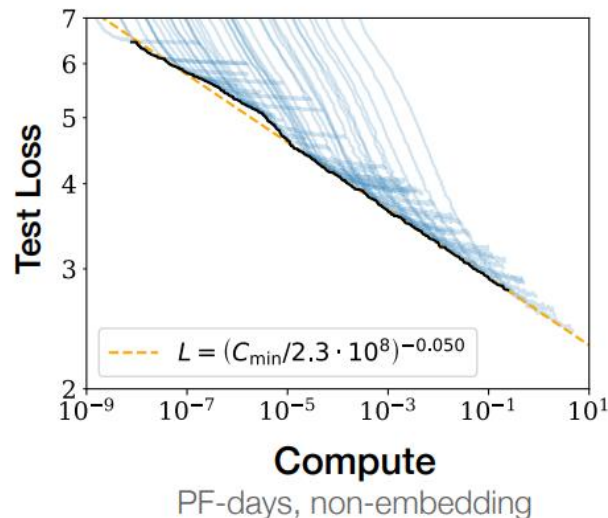
Most models are trained for approximately 300 billion tokens

**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, the dataset size, and the 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.

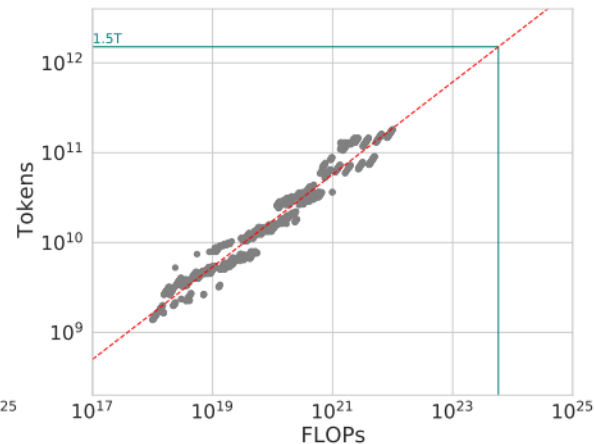
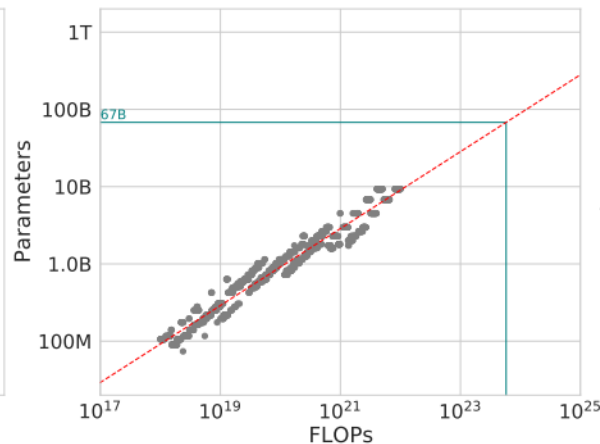
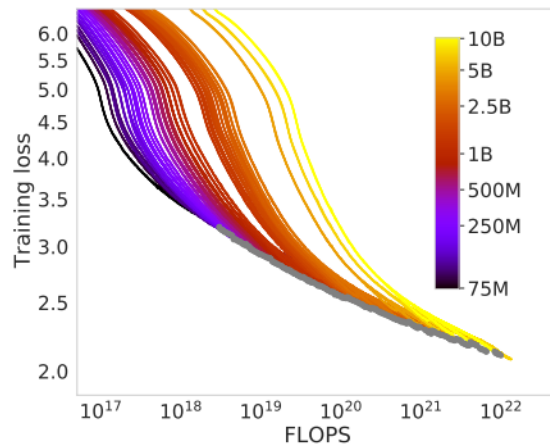


# Trends in model scaling

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

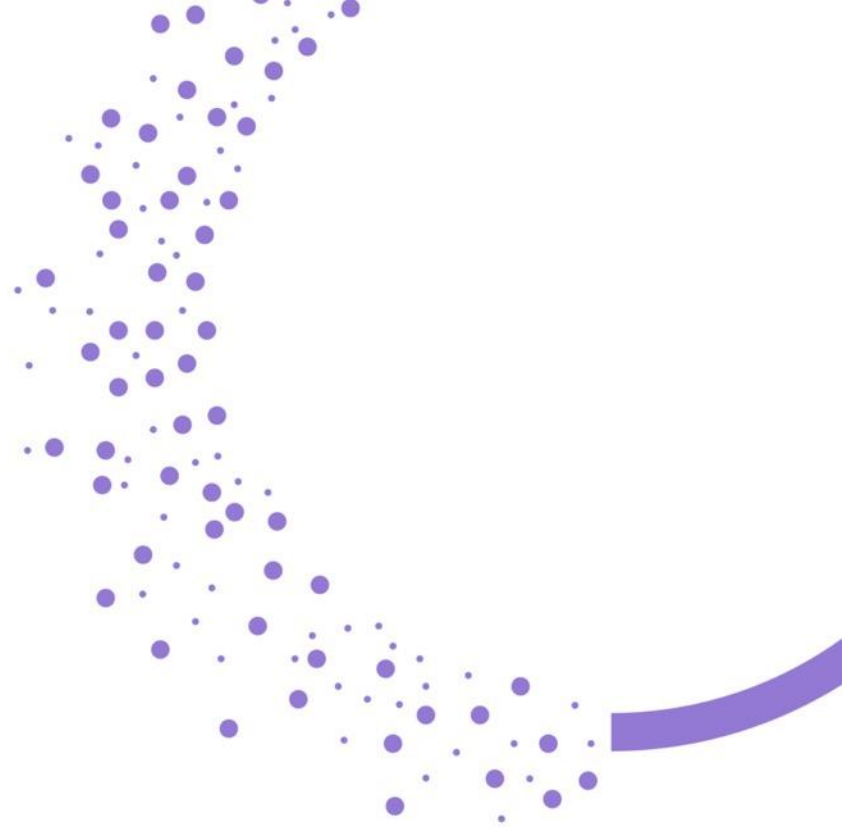
Jordan Hoffmann\*, Sebastian Borgeaud\*, Arthur Mensch\*, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, 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 undertrained, 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-

# NLP over the last years

Artificial intelligence  
Research Center

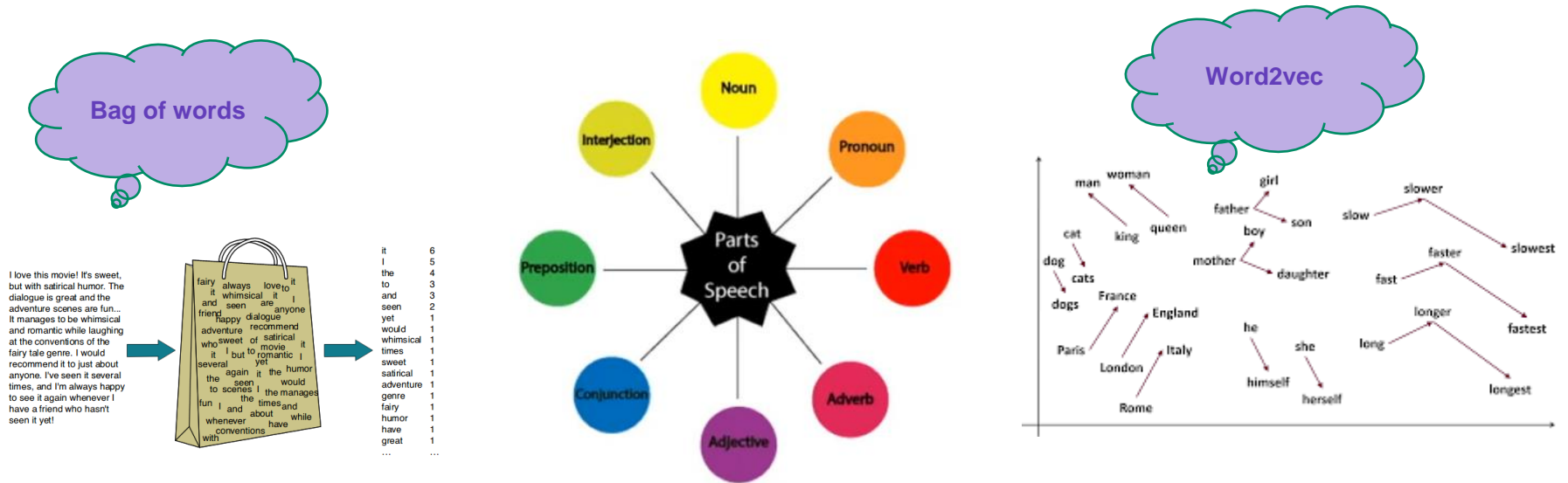
TII – Technology Innovation Institute





# NLP in the past

## Statistical NLP

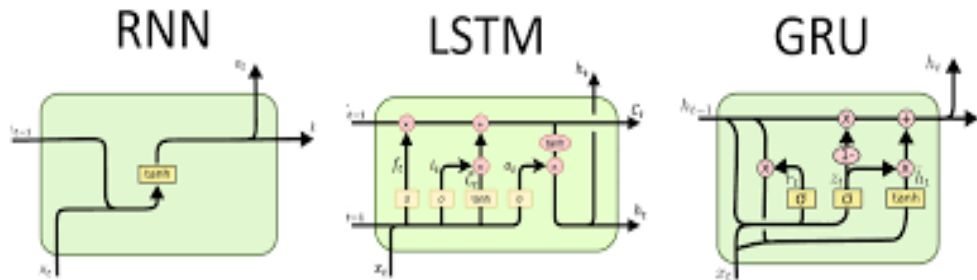
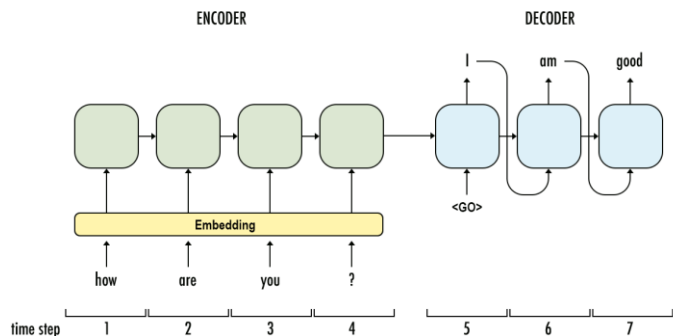


Most of natural language processing systems were based on simple statistical rules or non-complex Machine learning algorithms. The capabilities of these systems were limited to few tasks.

# 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 Toronto

### 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

Jacob Uszkoreit\*  
Google Research  
usz2@google.com

Llion Jones\*  
Google Research  
llion@google.com

Aidan N. Gomez\*<sup>†</sup>  
University of Toronto  
aidan@cs.toronto.edu

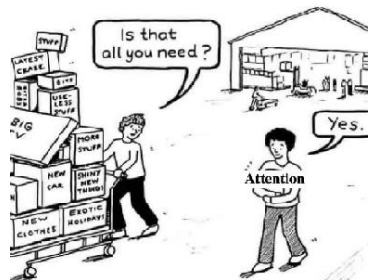
Lukasz Kaiser\*  
Google Brain  
lukaszkaiser@google.com

Illia Polosukhin\*<sup>†</sup>  
illia.polosukhin@gmail.com

#### Abstract

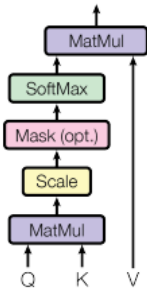
The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring 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

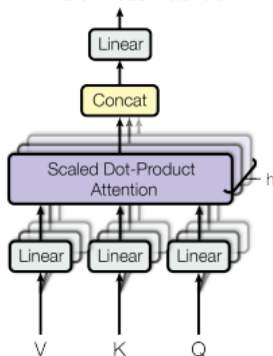


### Attention mechanism

#### Scaled Dot-Product Attention



#### Multi-Head Attention

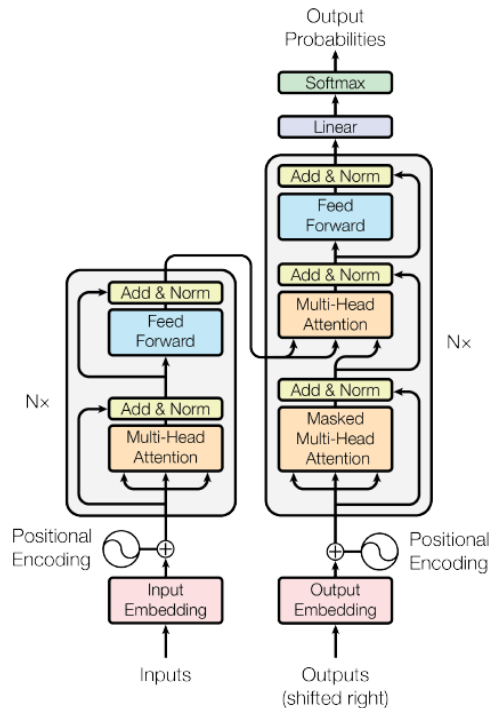


$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$



- Can learn Long dependencies
- Parallelizable

### Transformers



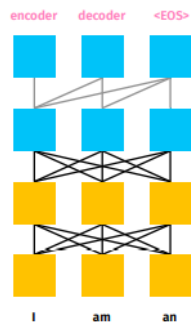
# Natural Language Processing: the Age of Transformers

Attention mechanisms revolutionized the way we do NLP

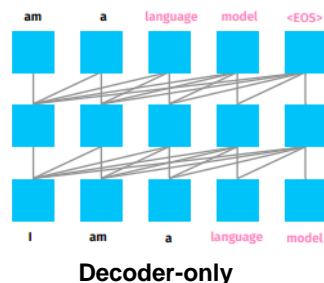
SOTA NLP models today are composed of a set of stacked multi-head attentions (transformers) : Encoder-based, Decoder-based or Encoder-Decoder models.

Examples of transformers-based language models :

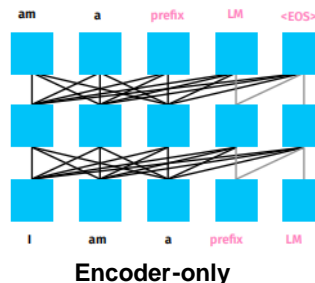
**encoder-decoder**  
e.g. T5



**autoregressive LM**  
e.g. GPT



**prefix LM**  
e.g. BERT



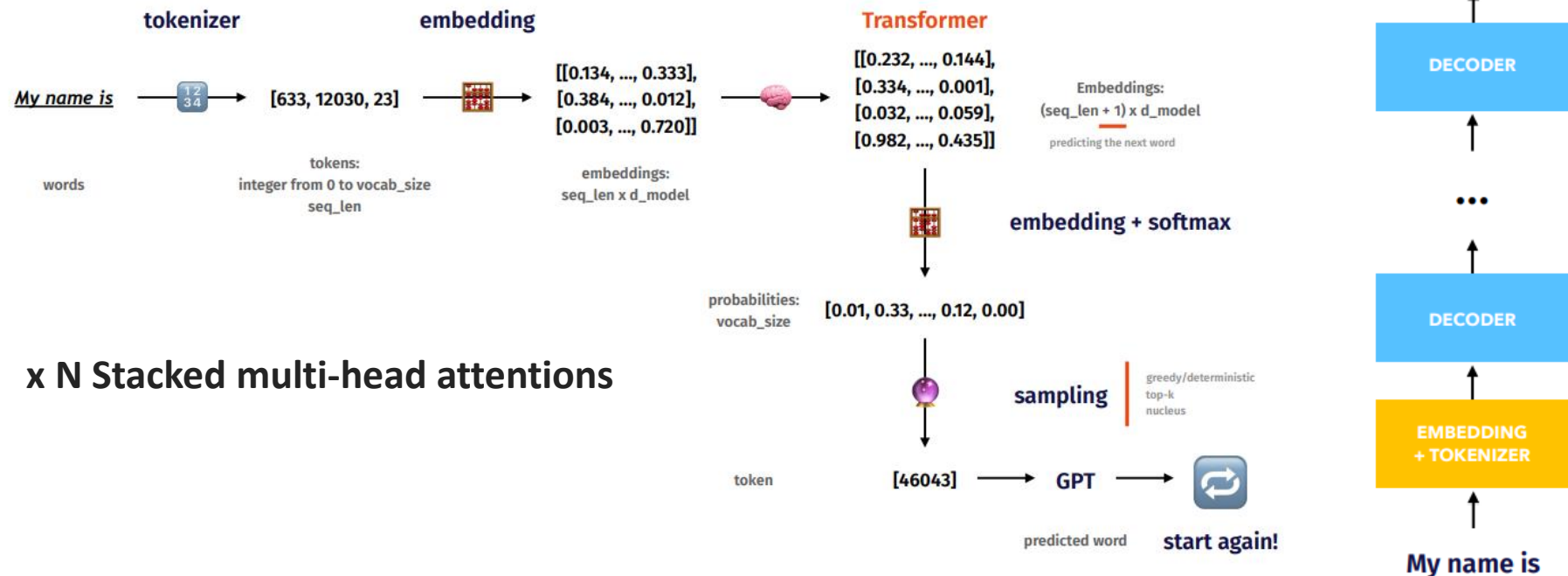
- Self-supervised models : MLM, next word prediction, sentences order ...
- Parallelizable with teacher forcing

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



# Focus on Decoder-only architectures ( E.g. GPT models )

## Predicting the next token

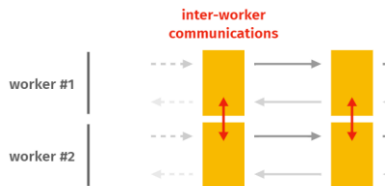


x N Stacked multi-head attentions

# Large and deep attention-based neural architectures

## Parallelism in the scope

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

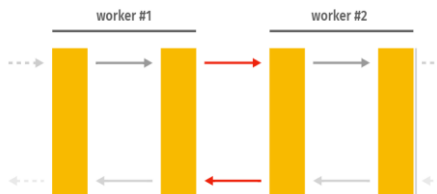


Simple to implement and to scale

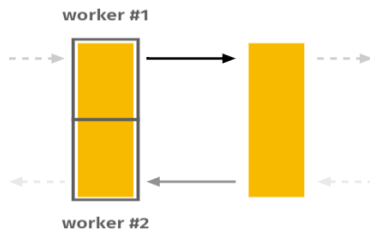


Memory

- **Pipeline Parallelism** : Horizontal parallelization over the layers



- **Tensor Parallelism** : Parallelism at the layer level



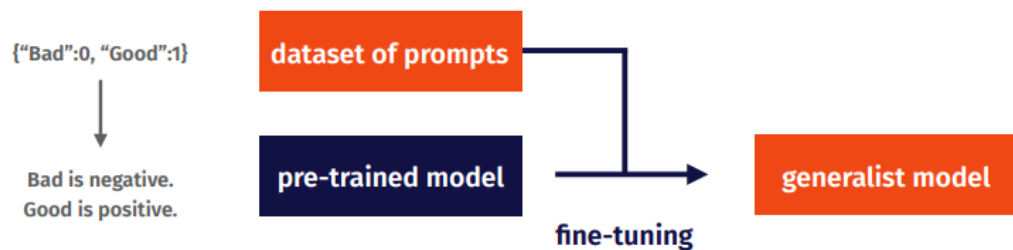
Requires Algebraic operations at the layer level

# LLM : Multi-task models

At scale, unique capabilities arise few-shot learning

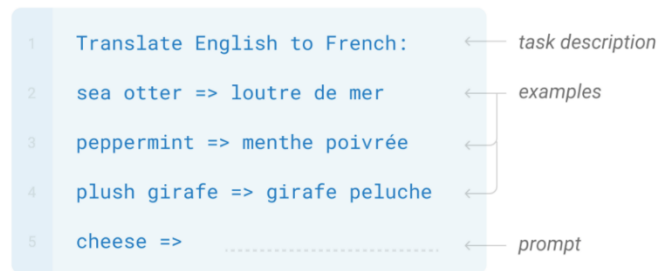
Power of extra-scale language models, **No need to fine-tuning**: Larger models can deal with unseen tasks on-the-fly

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

# Noor

## An Extra-Large Arabic Generative Model



# Noor Model

## Largest language model for Arabic in the world

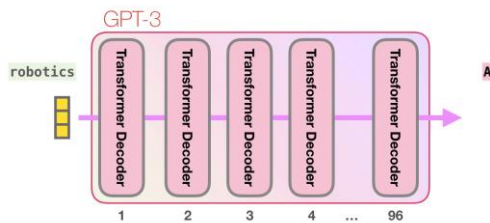
The Noor project expands upon the existing Arabic models, introducing 1.3B, 2.7B, 6.7B and 13B models. Therefore, Noor-13B is the largest NLP model for Arabic in the world.



## SOTA Arabic models

- AraBERT
- hULMonA
- ARAGPT-2: 1.47B parameters
- AraT5

## Similar architecture to GPT-3



- Noor is trained in self-supervision fashion to predict the next token
- Diversified sources of text: News, Government, Poetry, Crawl
- One epoch for training



Arabic is encoded in two bytes

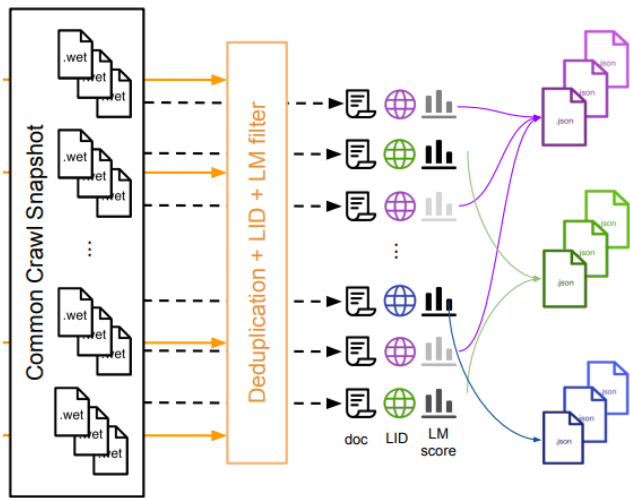


Source of data	Quality	Number of tokens [GTokens]	Duplication	Total tokens
Common Crawl	Low	44	1	44
C4	Low	19	1	19
ArabWeb16	low	7	1	7
OpenSubtitles	Medium	0.2	2	0.4
Wikipedia	High	0.3	3	1.2
News	High	5	2	10
Books	High	0.1	4	0.4
Pretraining datasets	High	0.6	2	1.2
<b>Total</b>	-	76	-	83

# Noor Model

## Processing the data

- Removing the diacritics
- Replacing some characters:  $\bar{\text{ا}} \bar{\text{ا}} \bar{\text{ا}}$  by  $\text{ا}$
- [CCNet](#): Extracting High Quality Monolingual Datasets from Web Crawl Data

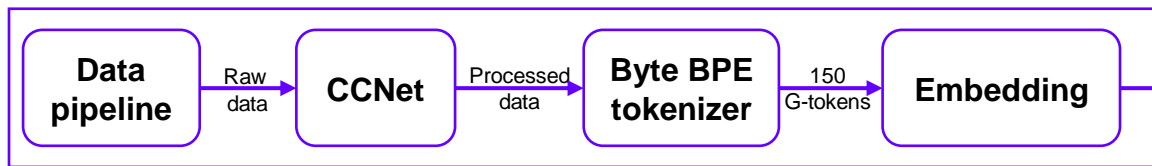
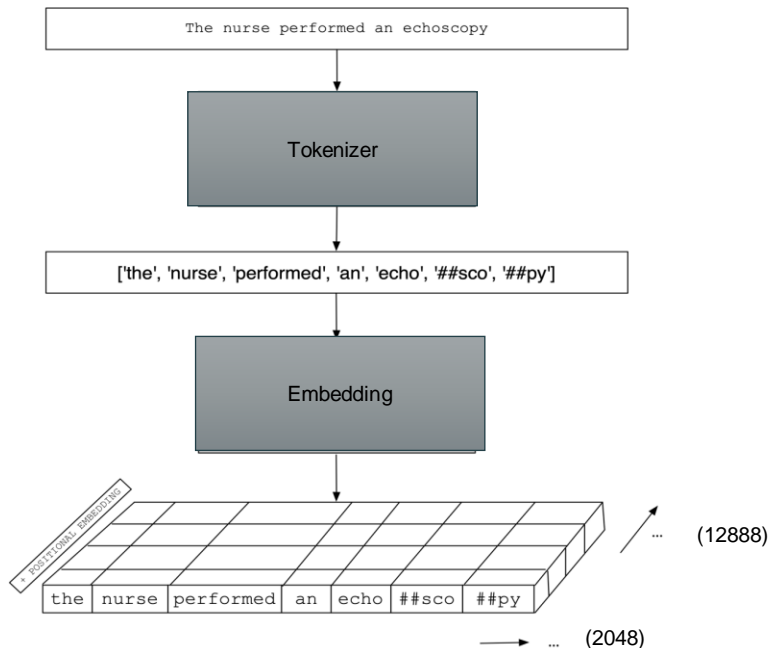


- **Deduplication** : Removing the duplicates at the paragraph/text level
- **Language Identification** : Identify the language of the text for every document
- **LM filtering** : Classify every text according to its perplexity score into head( high quality), middle, and tail( low quality )

# Noor Model

## Tokenization

- Three candidates :
  - ❖ **BPE at Byte level: Retained**
  - ❖ BPE with sentencePiece
  - ❖ Morphological tokenizer: Inefficient in Inference ( not adapted to production tasks )
- Perfect coverage rate with a good compression factor
- Vocabulary size : 50257 tokens
- Embedding dim : 12888
- Max tokens : 2048



# Noor Model

## Training

1.6M Batch size  
13B parameters

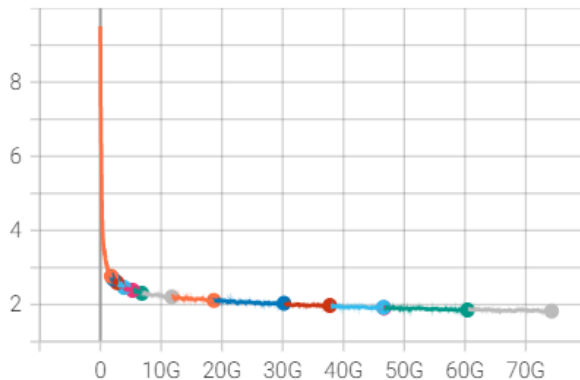
Multi-head attention  
layers

HPC of 160 A100

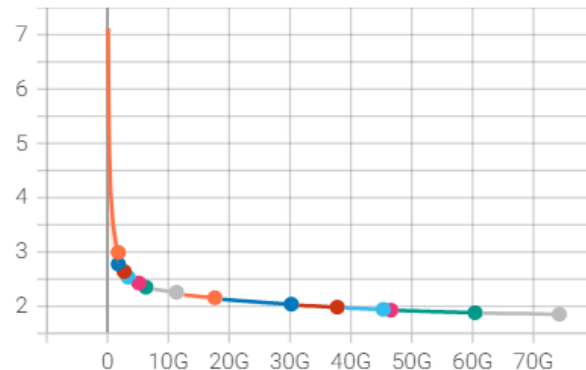
Cross Entropy loss  
Learning rate with a cosine schedule



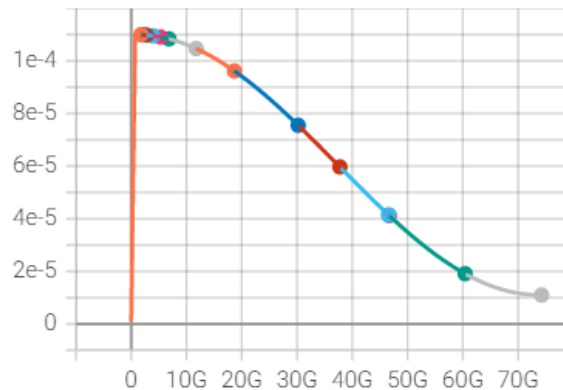
train\_loss  
tag: train\_loss



val\_loss  
tag: val\_loss



lr  
tag: lr



# Noor Model

## Inference

- Noor Model can fit in the memory of an A100 GPU for inference
- Multi-task model : Prompts + Few-shot

### Sentiment Analysis

```
prompts_few_shot = """
*قضيت يوماً جميلاً جداً : ايجابي
* تخرجت لمحاولة سرق : سلبي
* قمت بالاستمتاع بجمال الطبيعة : ايجابي
* الجو جد جميل اليوم : ايجابي
* نلقينا وقتاً ممتعاً شديداً : سلبي
* حصلت على نقطة ممتازة : ايجابي
"""
```

```
سليبي : * اصعباً فرصة الالتحاق بالمرتبة الأولى :
ايجابي : * احتل الصدارة عن جدارة و استحقاق :
ايجابي : * تم تكريمه اعترافاً بجماله الجيد :
ايجابي : * تحفلت بعيد المولد النبوي :
سليبي : * أشاد بمزودده امام الملء :
ايجابي : * عقدا قرانها في جو عائلي :
سليبي : * فاك فيروس كورونا بالحد من كبار السن :
سليبي : * فقد العديد من الإبراء ارباحهم نتيجة حادثة سير خطيرة :
ايجابي : * حيا الملك الاسباني بمخاضه تأجيله :
```

### Question/Answering


```
prompts_few_shot = """
. لتأكي الثورة الفرنسية في عهد لويس السادس عشر، الذي ساهم في عام 1792م، بتشكيل الصورة الأولية للجمهورية الفرنسية
رقم 22، من لائحة عدد السكان على مستوى العالم، كما أن متوسط أعمار سكانها 41.4 سنة. رئيس حكومتها هو كاستيكن
ارتباطها بالفن والأدب والثقافة، وارتباطها بالرفاهية والترفيه، مما يجعلها ضمن أكثر المدن التي تحظى بمعدل الجذب السياحي
اجب عن الأسئلة التالية
. سؤال : اين تقع فرنسا ؟ جواب : في القارة الأوروبية
. سؤال : كم عدد سكان فرنسا ؟ جواب : 65.424.082 نسمة
. سؤال : كم متوسط عمر سكان فرنسا ؟ جواب : 41.4 سنة
. سؤال : متى سكنت القبائل الرومانية فرنسا ؟ جواب : في فترة الخمسينيات مابعد الميلاد
"""
```

```
سؤال : ما هو نظام الحكم في فرنسا ؟ جواب : نظام جمهوري برلماني
سؤال : ما هي مرتبة فرنسا من حيث عدد السكان ؟ جواب : في المرتبة رقم 22
سؤال : ما هي المنظمات التي تنتمي اليها فرنسا ؟ جواب : منظمة الامم المتحدة
سؤال : ما هي لغة فرنسا الرسمية ؟ جواب : الفرنسية
سؤال : ما هي العملة المتداولة في فرنسا ؟ جواب : اليورو
سؤال : لماذا يفضل السياح الاجانب باريس ؟ جواب : لانها مدينة جميلة
سؤال : من هو رئيس حكومة فرنسا ؟ جواب : كاستيكن
سؤال : كم امتداد نهر السين في باريس ؟ جواب : كيلومترا 1.3
```

### Generate a text from a title

```
text = """
عنوان : فرنسا دولة بتقافة عريقة
رلسا العديد من الاشياء الدارزة التي دعمت الإنسانية في مجالات العلوم والأدب والفن والمعارف الأخرى
عنوان : تكتب الأوزون يشكل خطرا على الأرض
بة حيث يؤدي ذلك إلى خلل في جهاز مناعة الإنسان والإضرار بالحيون والنباتات والارتفاع الإشعاعية بمرطبان الجلد
ن العسر بالمحصور وبالتالي انخفاض القدرة الإنتاجية مما يهدد الأمن الغذائي. على سطح الكرة الأرضية
من وأما الأجزاء المائية الأخرى فيعتمد العلماء بأنها أكثر أمنا من غيرها فكمية وجود الماء التي جميعها
مية مسارة في مناخ الأرض علما بأن مركبات الكلوروفلوروكربون هي ضمن غازات الاحتباس الحراري
عنوان : كرة القدم رياضة شعبية
ة، المألوفة في جميع الأوجال ومن جميع البلدان هي التي تجعل كرة القدم أكبر رياضة على هذا الكوكب
عنوان : كورونا فيروس خطير
تتعدد لدرجة تستلزم دخولهم المستشفى، وقد يصل بهم الحال إلى الاعتماد على أجهزة التنفس الاصطناعي
ة، أو ناء السكري. ويشبه ذلك ما يحدث عند الإصابة بأمراض الجهاز التنفسي الأخرى، مثل الإنفلونزا
عنوان : الحفاف يهدد الأرض
مقال : *مختبر الحفاف
"""
```

```
لبنها خطرا على الأرض، حيث يسبب في موت الكثير من الكائنات الحية، ويهدد حياة الملايين
ويهدد الحفاف اذ اعم لنبات نحر الطاع، حيث له يحث علما تلى كمية الهواء الممتعة على سطح الأرض
ويهدد الحفاف أيضا العديد من أنواع النباتات والحيوانات، بما في ذلك الطيور والاسماك والنباتات، فضلا عن الحيوانات البرية
.ويؤدي الحفاف الى جفاف المحاصيل الزراعية، مما يؤدي الى حداث نقص في الغذاء، ويؤدي الى نقص في المواد الغذائية، مما يؤدي الى المجاعة
. ويؤدي الحفاف أيضا الى زيادة معدل هطول الأمطار، مما يؤدي الى زيادة الضغط على موارد المياه العذبة، وزيادة تلوث المياه
.ويؤدي الحفاف أيضا الى زيادة مخاطر انتشار الأمراض، مثل الملاريا
.ويؤدي الحفاف أيضا الى زيادة المخاطر الصحية النشئة بتلوث الهواء
ويؤدي الحفاف أيضا الى ارتفاع معدلات البقران، مثل سرطان الثدي، وسرطان القولون، وسرطان الرئة، وسرطان البنكرياس، وسرطان المثانة
.رئة، وسرطان الكلى، وسرطان المثانة، وسرطان البروستاتا
```

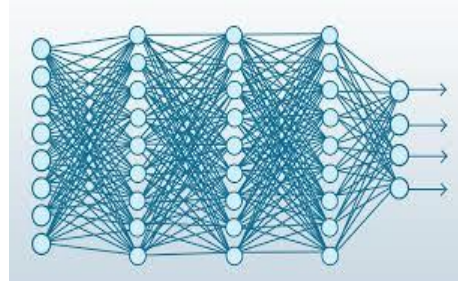
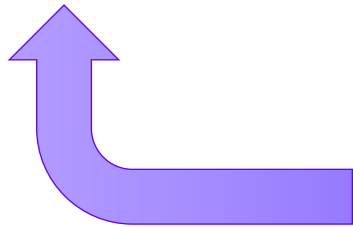


A Holistic Assessment of the Carbon Footprint of  
Noor,  
**a Very Large Arabic Language Model**

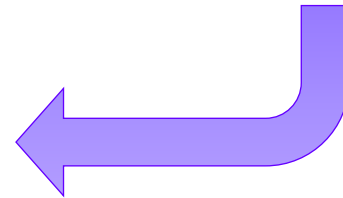
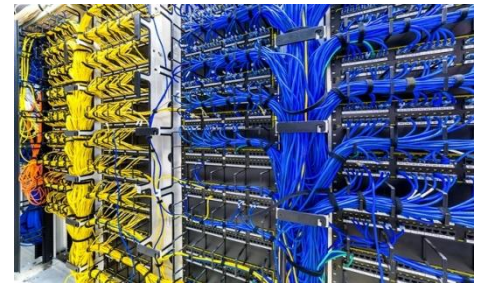
Artificial intelligence  
Research Center

TII – Technology Innovation Institute

# AI and Environment

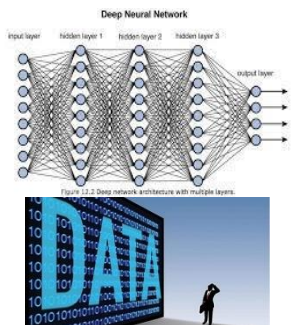


- The major progress in AI has been done through deep learning
- The demand for compute for deep learning is increasing exponentially
- Is deep learning a brute force technic ?



# Energy consumption of extra-scale models and their CO2 footprint

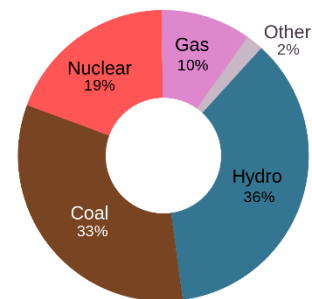
## Factors



CPU

GPU

TPU



### The model size and the size of the dataset:

Given the model size and the data size, we can determine the number of FLOP required to train the model.

During serving, the model size determines the number of FLOP per forward pass.

### The hardware characteristics:

The number of FLOPs the hardware can perform, and its nominal power ( for instance  $A100 \neq V100$  )

### The efficiency of the datacenter:

Energy required to cool down the datacenter and other needs

PUE : Power Usage Effectiveness allows to measure the energetic efficiency of the datacenter

### The energy supply mix :

Energy sources powering the datacenter

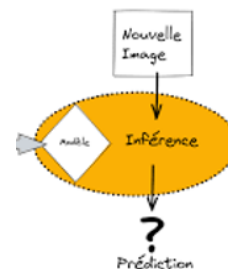
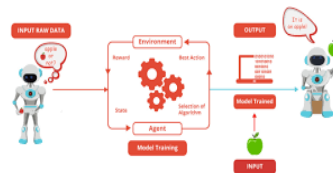
May depend on the region of the installation

CO2 emission per kWh of energy consumption



# Energy consumption of extra-scale models and their CO2 footprint

## Sources



### Storage and Transfer costs

- Data is stored on servers: Average Peak Power per TB 11.3W/TB
- Transferring data between nodes requires energy as well : average of 6.38 kWh/TB transferred.

### Experimentations

- Data curation and processing
- Research and development for data validation, tokenization and hyperparameters' finetuning

### Training

- Training of the final versions of the model

### Inference


- Serving of the model
- How many tokens are generated by the model ?

### Other expenses

- Personal laptops
- International collaboration
- Emails and Video chat
- Travels and commutes

# Storage and Transfer costs

## Storage

- We used the services of Amazon S3  to store our data
- Power per TB : 11.3W (PUE of 1.6 and a redundancy factor of 2 for backup)
- Assumption : Data is stored for 6 months; the duration of the project
- The server is located in Bahrein : 1,2 kgCO2/kWh

Data	Storage in TB	W consumption (kWh)	CO2e (Kg)
Curated	2	99	118
Bulk	25 TB for 6 months 210 TB for 1 day	1300	1544
Model	5.7	300	356

- 1.7 MWh** of energy consumed in Storage
- 2 tons** of CO2e

## Transfer :

- In average, 6.38 kWh per TB transferred

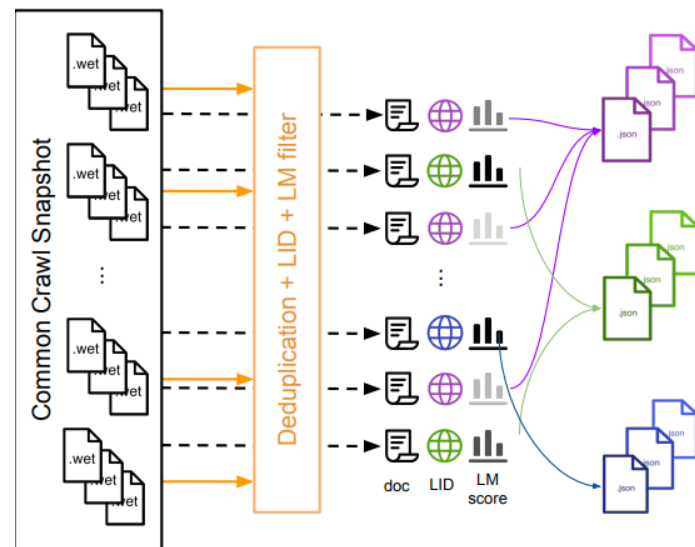


Data	TB	Transfer	Transfer coeff
CC	210	Downloaded on the preprocessing server once	x1
Processed CC	25	Moved once to our archival machines and another time to the HPC used for training	x2
Curated data	2	downloaded once, moved to the archival machines, and then moved to the HPC	x3
Model	5.7	Once moved from the HPC to the archival machine and once to the inference servers	x2

- 1.8 MWh** of energy consumed in Transfer
- 2.1 tons** of CO2e

# Data processing

- We process text data with **CCNet**
- **CCNet** is a pipeline in charge of deduplication, language identification, and language filtering
- We use a CPU cluster of 768 cores, split to 16 nodes ( Google Cloud Platform )
- The cluster is based in Netherlands : **410 gCO2e/kWh**
- We estimate the average power consumption of each node to be about 350W  
→ The power of the cluster : 5.6 kW
- 21 dumps of CC + curated data
- 381 wall-clock hours
- PUE of 1.1
- **2.35 MWh** of energy consumed in Data processing
- **0.96 tons** of CO2

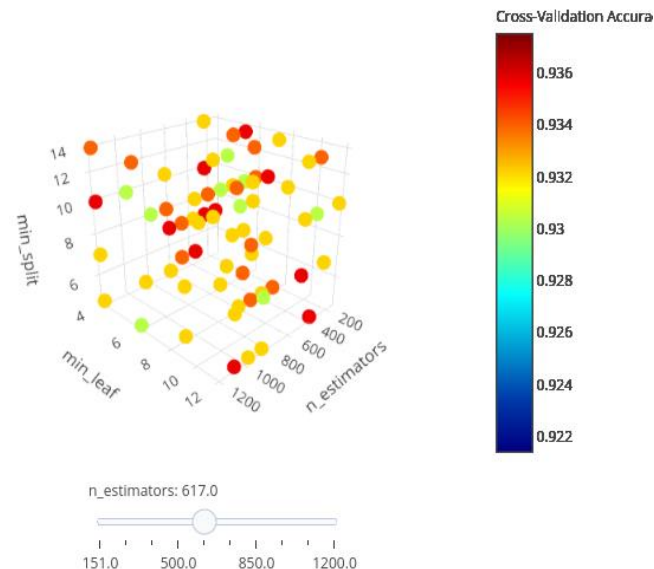


# Experimentations

## Research & Development

- 2 Tokenizer candidates : We train two small models of 350M parameters
- Tuning the hyperparameters
- Establish scaling laws
- We used Meluxina super-computer for R&D experiments
- The HPC is located in Luxembourg ( PUE = 1.35 and 60g of CO<sub>2</sub>e per kWh)
- Each node of Meluxina is made of 4 A100 SXM 40GB with a TDP of 400W, and two AMD EPYC 7763 CPUs with a TDP of 280W
- We estimate the total compute spent in this phase to be **16,800 A100-hours**
- **10.7 MWh** of energy consumed
- **0.65 tons** of CO<sub>2</sub>

Cross-Validation Accuracy Varying Hyperparameters



# Training

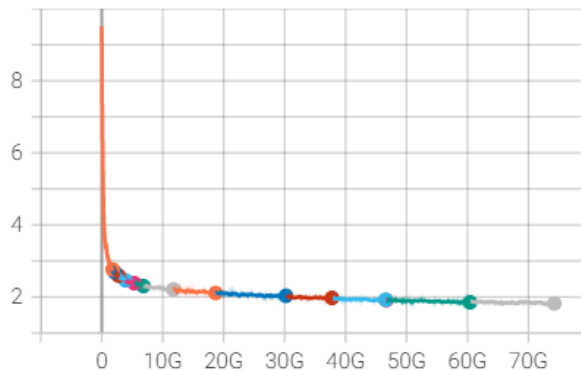
- Training compute requirement in FLOP :  $C = 6 \times N \times D$  (  $N$  : number of tokens,  $D$  number of parameters )
- Using the previous formula, we can approximate the training budget, energy consumption and eCO2.
- Effective observed throughput : 100 TFLOPs per GPU
- Noor-HPC : 20 nodes, each contains 8 A100 80GB and 2 AMD EPYC 7763 CPUs
- PUE : 1.5 and 600 gCO2e per kWh

Table 1: Training compute budget and energy used for training the Noor models. Assuming a pretraining dataset of 150B tokens and a throughput of 100 TFLOPs per A100.

Model	Budget [PF-days]	Budget [A100-hours]	HPC	Consumption [MWh]	Footprint [tCO2e]
1.3B	13.5	3300	MeluXina	2.1	0.13
2.7B	28.1	6800	Noor-HPC	4.8	2.9
6.7B	69.8	17000	Noor-HPC	11.8	7.1
13B	135	33000	Noor-HPC	22.9	13.8

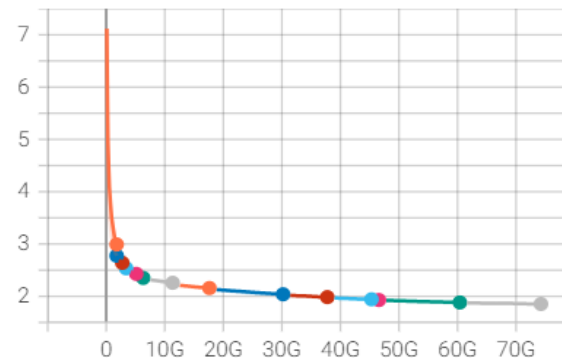
train\_loss

tag: train\_loss



val\_loss

tag: val\_loss



# Inference

## A prospective estimate of Inference cost under strong assumptions

- An A100 GPU is enough to hold Noor model of 13B parameters
- Two assumptions :
  - Inference time per generated token is constant (~50ms), whichever the number of processed tokens (up to 512 processed tokens roughly)
  - Batch size is 1

→ An A100 can generate up to 72000 tokens per hour

- 400W for the A100, 70W for the CPU, and PUE = 1,1

→ **26 Joules** of W consumption per generated token

→ **3 billion tokens would have to be generated for inference costs to catch up with training costs**

- Assume the world average emission per kWh (475 gCO<sub>2</sub>e/kWh)

→ **300.000 tokens generated per Kg of CO<sub>2</sub>e**

**OpenAI reported to generate 4.5 billion words per day with GPT-3**



**Assuming the same number of generated tokens for Noor :**



**30 tons of CO<sub>2</sub>e per day for serving**

# Additional costs

In addition to the development costs, other factors contribute the energetic and environmental bills of Noor :

- Personal Laptops, emails and video-conferences: ~ 1 MWh (rough estimate)

→ 0.42 tCO<sub>2</sub>e

- International cooperation: 3 round-trip flights of four scientists between Paris and Abu Dhabi.

→ 6.4 tCO<sub>2</sub>e

# Summary

## Total Energy consumption and CO2 footprint (excluding inference)

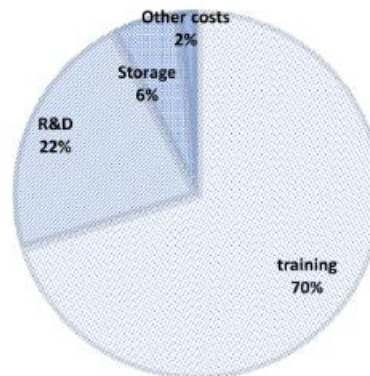
- Total 59.14 MWh of energy consumption of which 70% went for training
- Total 36.5t of CO2 emission, 65% went for training and 18% for international flights

### Main findings :

- Despite being substantial, the energetic and environmental bills are not only about the final trained model
- Inference can overtake easily the training cost (in one or few days of intensive serving)
- The breakdown of CO2 footprint is highly dependent on the localization of the workloads and the local carbon intensity of the electricity mix
- Some exogenous factors to development, such as international flights, can have a substantial carbon footprint

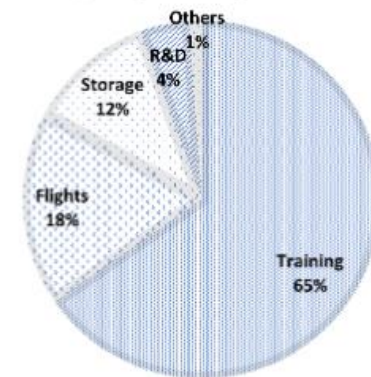
ENERGY CONSUMPTION IN MWH

training R&D Storage Other costs



CO2 EMISSION IN TONS

Training Flights Storage R&D Others





# Best practices and recommendations

## Modeling & Engineering

- **Efficient architecture:**
  - **Mixture of Expert MoE** : Splits the fully-connected layers of a transformer into distinct experts. This can bring significant energy saving since experts are only sparsely activated
  - Consider Sparsity in the network
  - Think of an optimal trade-off : data size – model size given a fixed compute budget ( *The new Paper of DeepMind that revisited Kaplan et al.* ) → Small models are less demanding to infer
- **Efficient inference :**
  - **Quantization**: reduces numerical precision at inference time and accelerates serving
  - **Distillation**: Training a smaller model from the outputs of a larger one
- **Efficient implementation :**
  - Achieve the best throughput possible per GPU
- **Think of new ways of doing Machine Learning :** No major discovery in the last years except increasing the size of neural networks

# Best practices and recommendations

## Hardware

- **Data Center choice**

  - Prioritize datacenters with high efficiency ( low PUE )

  - For instance, a PUE of 1.1 will decrease the energy consumption by 39% compared to the world average of 1.8

- **Local Carbon Intensity**

  - Carbon intensity of the electricity mix significantly impacts the final footprint

  - Locating training in an area with a clean mix is an easy step to take and that can drastically cut the project footprint

- **Efficient Inference**

  - Select a tailored accelerator for inference according to the model characteristics

# Best practices and recommendations

## Other practices

- **Minimizing exogenous impacts**

Minimizing high-intensity costs like international flights can reduce significantly the CO2 emission

- **Cost reporting and Offset**

The full cost to develop large deep learning models is rarely if ever reported in the literature

We highly recommend to the AI community to start reporting the full energy consumption and the CO2e of their projects

Consider the carbon footprint as a metric along with other performance metrics to evaluate the models



# Conclusion

## End-to-End Assessment of the energetic and environmental bills of Noor

- First exhaustive assessment of the environmental bill of a deep learning model
- Development cost of Noor :36.5 tons ( excluding production cost)
- To put this number in perspective : The average American emits 20 tons of CO2 a year and **a jet plane doing a roundtrip between San Francisco and New York has 180t of CO2 emissions**
- The main driver of CO2 footprint is the carbon intensity of the mix used to power the hardware ( *For instance, running all computations in France would have reduced the total footprint to 14.9 tCO2e, 42% of which from the international flights*)
- We highly encourage the AI community to consider the datacenter efficiency and its supply mix and to report the full CO2 bill of their projects
- Production : The energetic bill of large-scale inference is huge and can rapidly overtake the development cost → Raise awareness to adopt efficient inference practices



**Thank you**