



BUILDING LARGE-SCALE, LOCALIZED LANGUAGE MODELS: FROM DATA PREPARATION TO TRAINING AND DEPLOYMENT TO PRODUCTION

MERIEM BENDRIS | MIGUEL MARTÍNEZ

NATURAL LANGUAGE PROCESSING?

"Dentro de la inteligencia artificial, el procesamiento del lenguaje natural es un área donde los idiomas empiezan a destacar. Al entender el lenguaje de los textos impresos, los expertos pueden traducir de manera rápida y eficiente cualquier mensaje que sea necesario."

NATURAL LANGUAGE PROCESSING?

"Dentro de la inteligencia artificial, el procesamiento del lenguaje natural es un área donde los idiomas empiezan a destacar. Al entender el lenguaje de los textos impresos, los expertos pueden traducir de manera rápida y eficiente cualquier mensaje que sea necesario."

- Text entered by us
- Text generated by AI



ABOUT ME

Meriem Bendris



- Senior Deep Learning Data Scientist at NVIDIA
- Focus: Conversational AI and Natural Language Processing
 - PhD in Signal and Image Processing
 - Expertise in Machine Learning and large-scale Deep Learning
 - Riva PIC SA in EMEA

ABOUT ME

Miguel Martínez



- Senior Deep Learning Data Scientist at NVIDIA.
- Focus on Machine Learning, Natural Language Processing, Recommender Systems and Graphs.
- Expertise in GPU-Accelerating End-to-End Data Science Workflows.

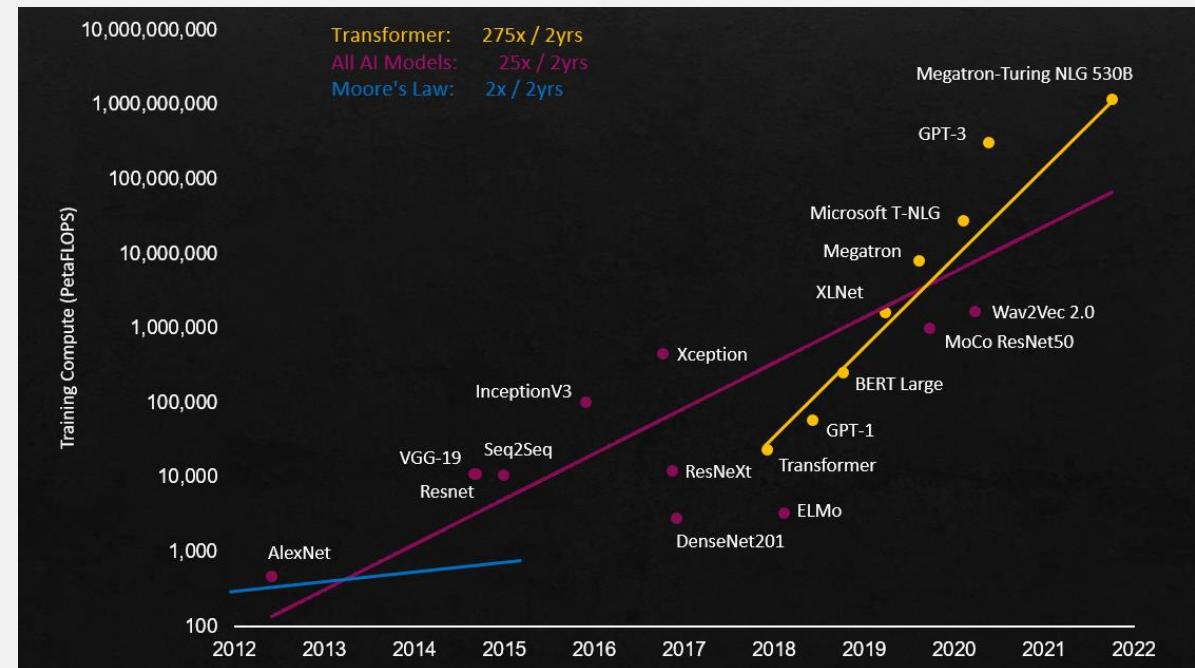
WHY LARGE MODELS

LARGE LANGUAGE MODELS

Large models, large datasets and large compute

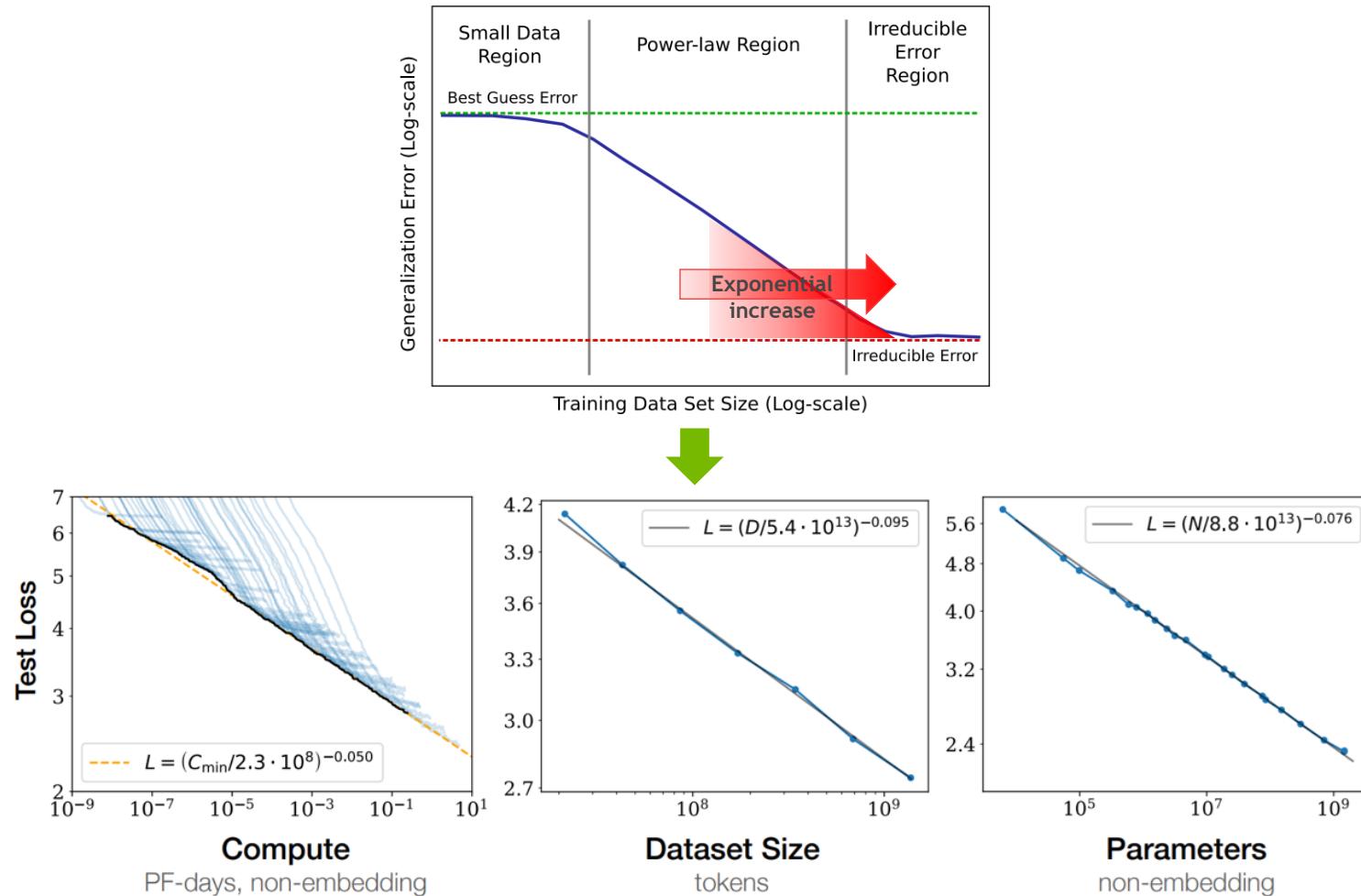
- Ability to efficiently collect and process large volumes of data
- Ability to efficiently train large models on large volumes of data
- Ability to cost effectively deploy large models

Exploding Complexity



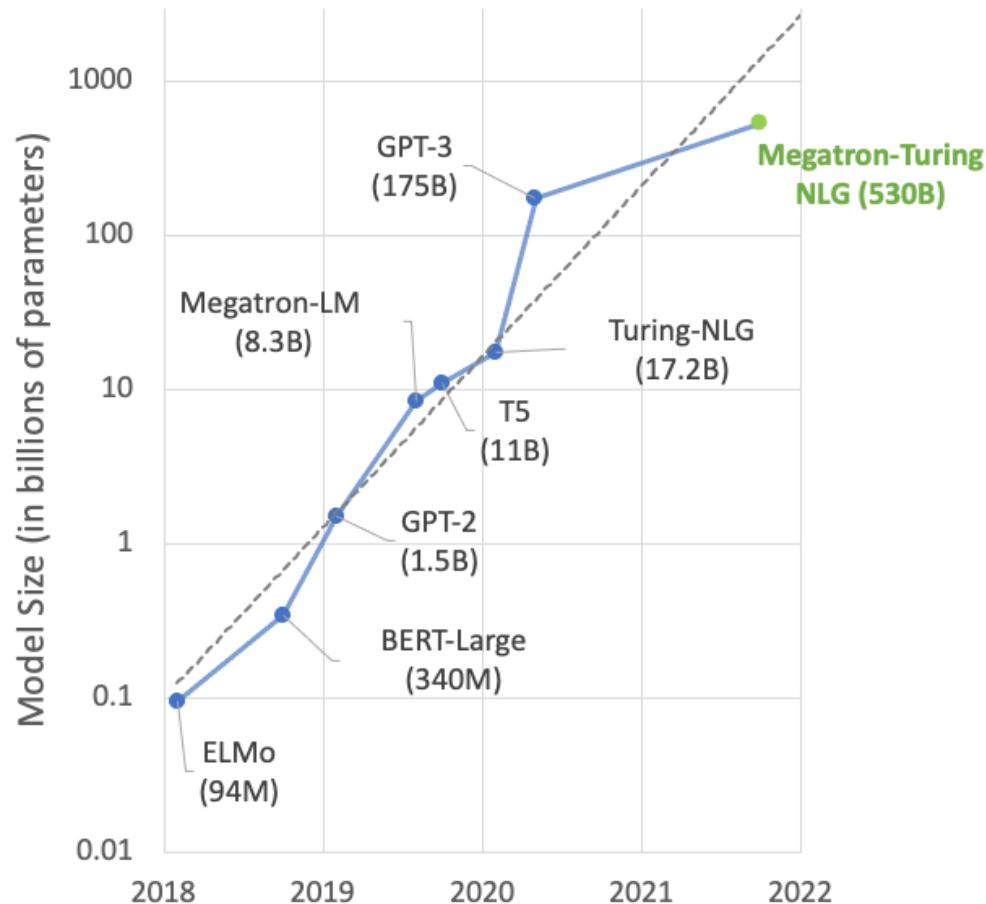
WHY LARGE NLP

The Scaling Laws



MEGATRON-TURING NLG 530B

The Trend Continues



WE TALK TO A LOT OF CUSTOMERS ABOUT NLP

Some EMEA countries we are working at



Germany



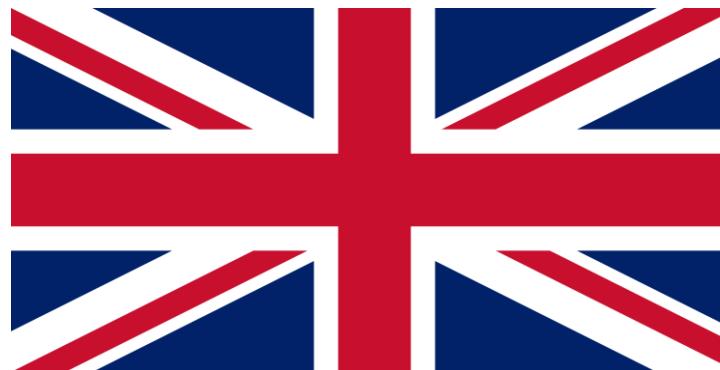
Spain



France



Israel



United Kingdom



Sweden

PERCEIVED AS A PROHIBITIVELY DIFFICULT PROBLEM

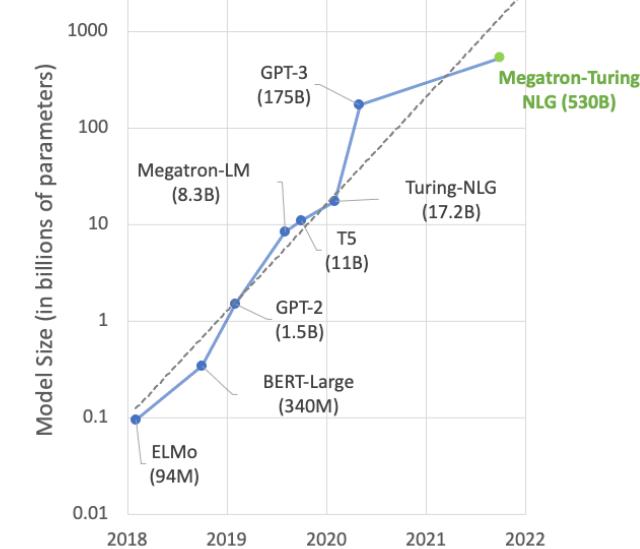
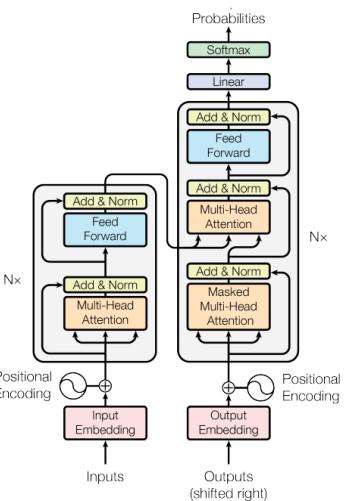


Figure 1: The Transformer - model architecture.



$$AGG(\{h_u^{(l-1)}, \forall u \in N_m(v)\}) = \frac{1}{|N_m(v)|} \sum_{u \in N_m(v)} h_u^{(l-1)}$$

$$\alpha_m = \frac{\exp(e_m)}{\sum_{m=1}^M \exp(e_m)}$$

$$h_v^{(l)} = \sum_{m=1}^M \alpha_m \cdot h_v^{(l)[m]}$$

$$h_v^{(l)[m]} = W^{(l)[m]} \cdot \text{CONCAT}\left(W_d^{(l)[m]} \cdot h_v^{(l-1)}, W_s^{(l)[m]} \cdot AGG(\{h_u^{(l-1)}, \forall u \in N_m(v)\})\right)$$

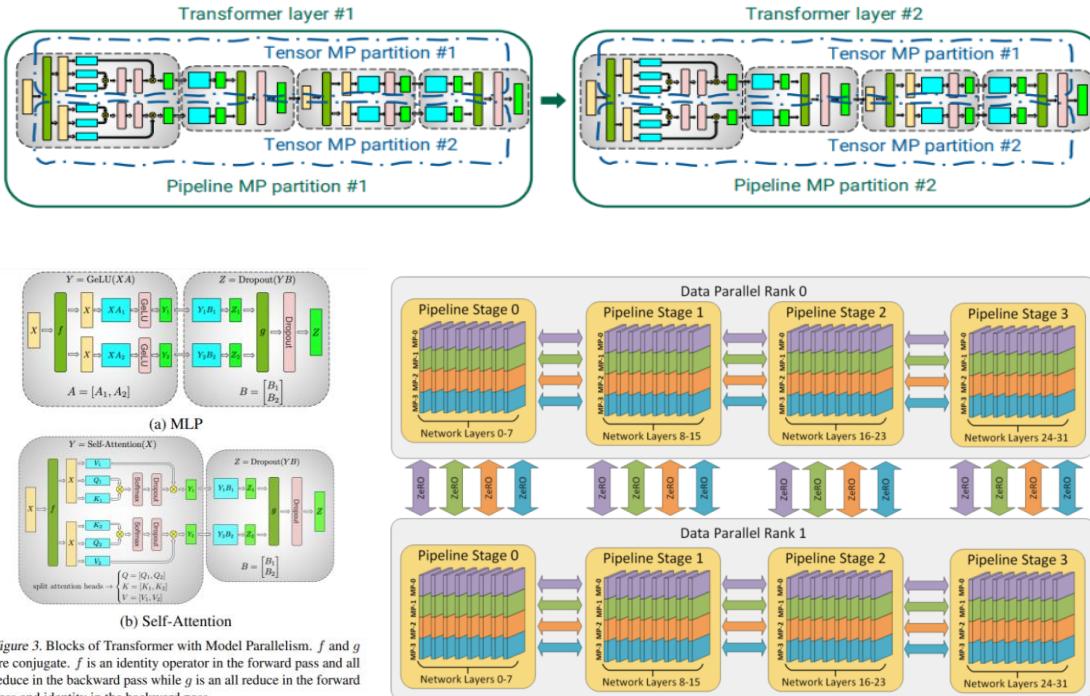
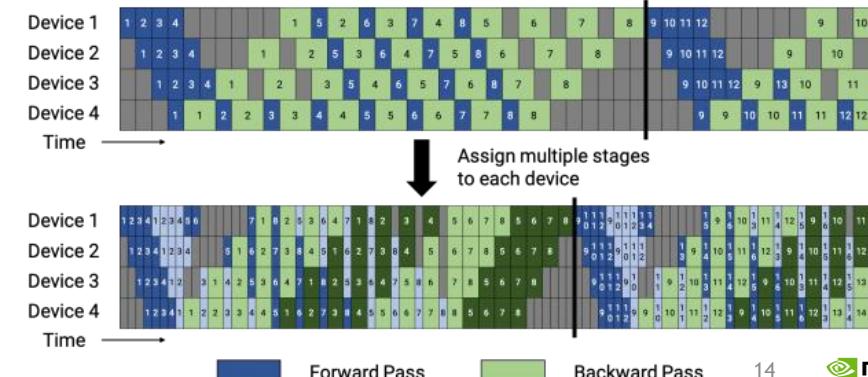
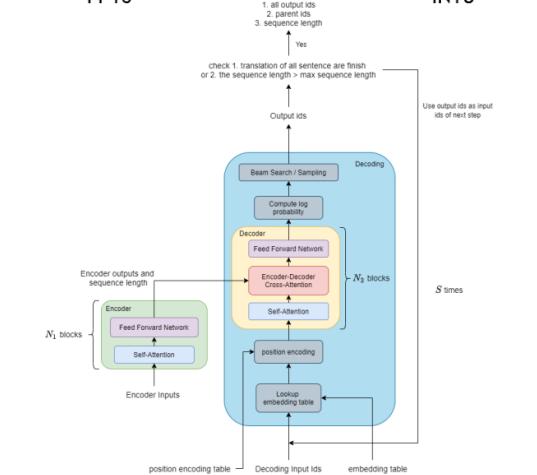


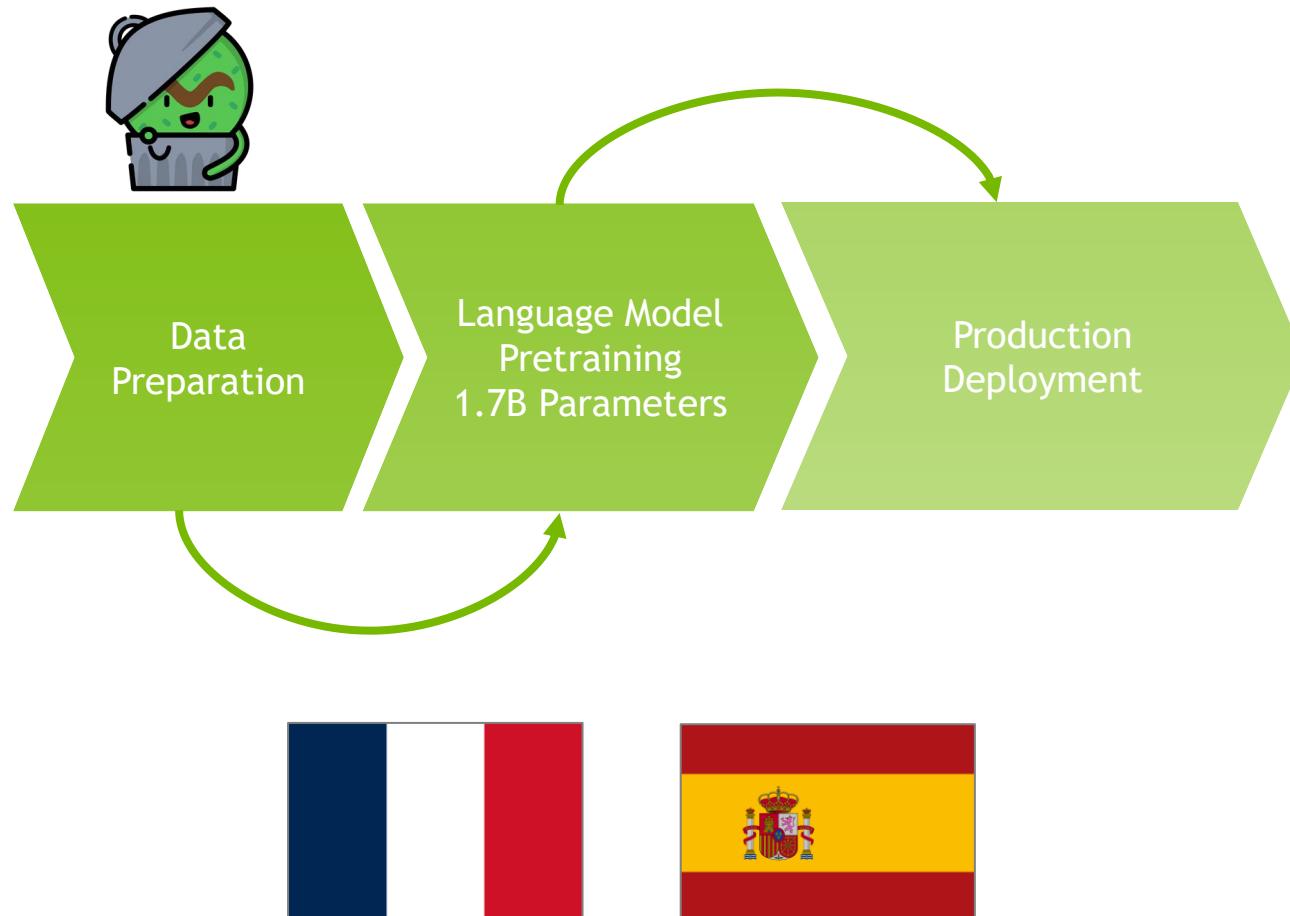
Figure 3. Blocks of Transformer with Model Parallelism. f and g are conjugate. f is an identity operator in the forward pass and all reduce in the backward pass while g is an all reduce in the forward pass and identity in the backward pass.

	FP16	INT8
0.34	64	76
3.75	134	119
5.64	217	21
1.12	4.7	0.68
2.7	1.43	-0.9

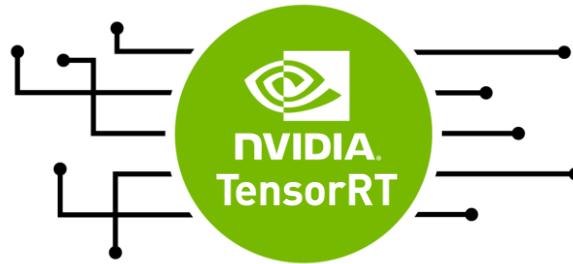


THE GOAL OF THIS WORK

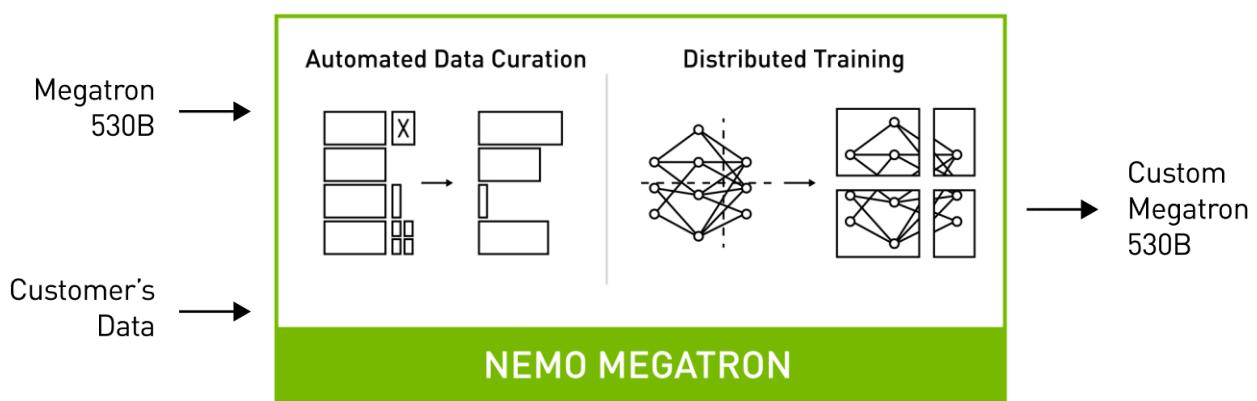
Our Journey



A WIDE VARIETY OF TOOLS



TRITON INFERENCE SERVER



NVIDIA/Megatron-LM

Ongoing research training transformer models at scale

NVIDIA/FasterTransformer

Transformer related optimization, including BERT, GPT

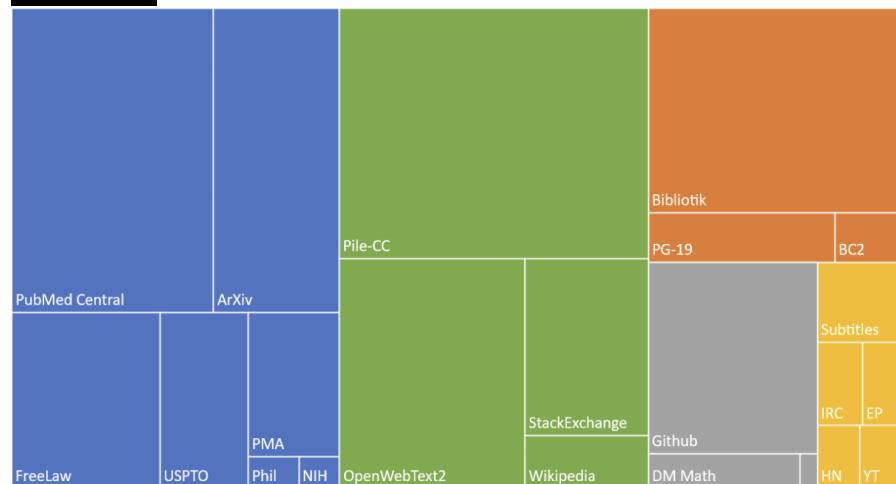


HUGGING FACE



SEVERAL PUBLICLY AVAILABLE DATASETS

The Pile



Common Crawl

Common Crawl
7 years of crawling
the internet



OSCAR

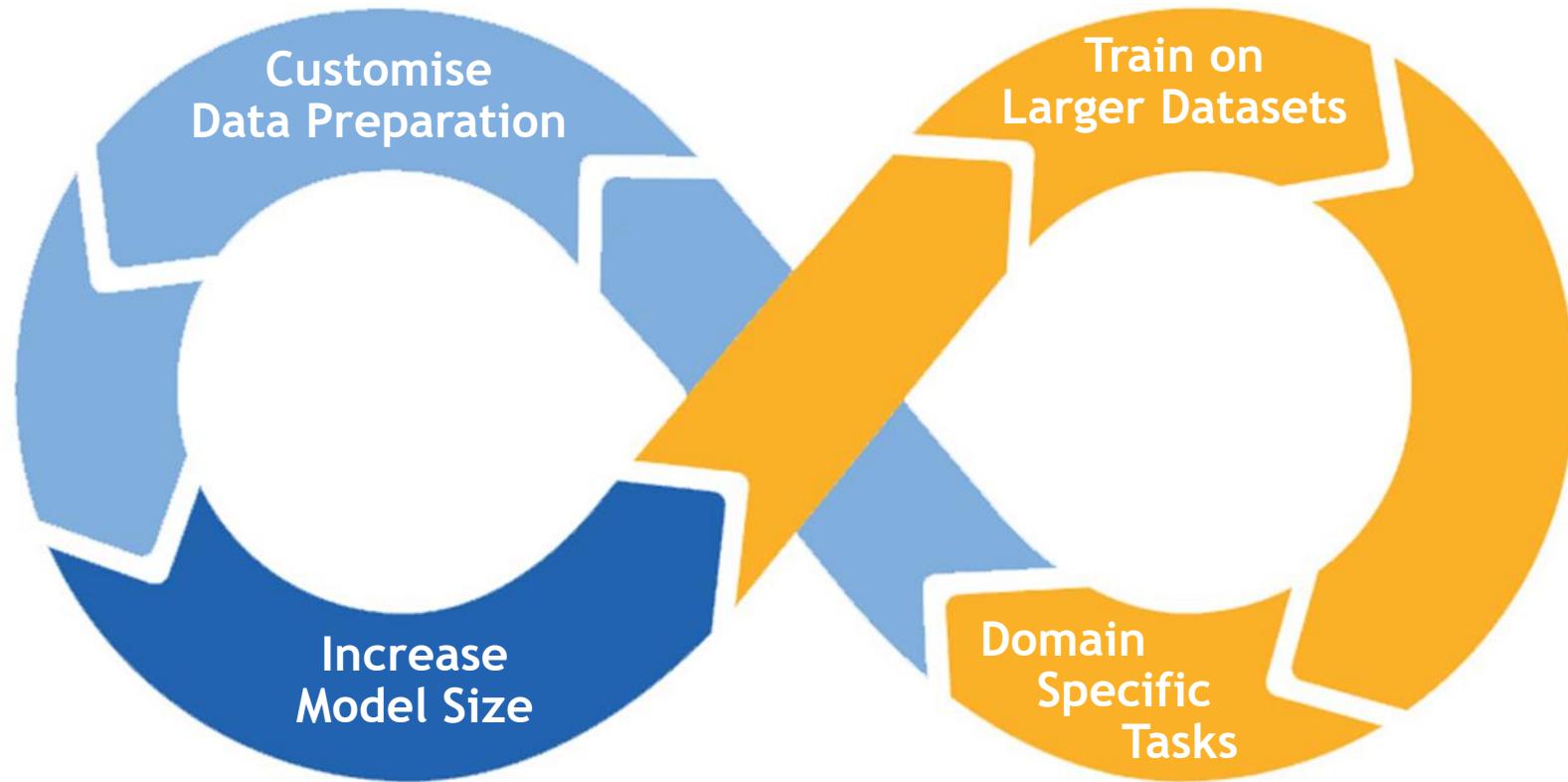
Open Super-large Crawled Aggregated Corpus
<https://oscar-corpus.com/>



A multilingual variant (101 languages)
of the [C4](#) dataset

<https://www.tensorflow.org/datasets/catalog/c4>
<https://paperswithcode.com/dataset/mc4>

DO IT ONCE, CONTINUOUSLY IMPROVE IT



- Hyperparameter optimization
- Adapt the model vocabulary to the customer use case
- ...

OUR 3 WEEKS JOURNEY

WEEK 1

DATA PREPARATION



DATA COLLECTION

OSCAR dataset



OSCAR

Open Super-large Crawled Aggregated corpus
<https://oscar-corpus.com/>

Language	Size		Words		Language	Size		Words	
	Orig	Dedup	Orig	Dedup		Orig	Dedup	Orig	Dedup
Afrikaans	241M	163M	43,482,801	29,533,437	Lower Sorbian	13K	7.1K	1,787	966
Albanian	2.3G	1.2G	374,196,110	186,856,699	Luxembourgish	29M	21M	4,403,577	3,087,650
Amharic	360M	206M	28,301,601	16,086,628	Macedonian	2.1G	1.2G	189,289,873	102,849,595
Arabic	82G	32G	8,117,162,828	3,171,221,354	Maithili	317K	11K	69,161	874
Aragonese	1.3M	801K	52,896	45,669	Malagasy	21M	13M	3,068,360	1,872,044
Armenian	3.7G	1.5G	273,919,388	110,196,043	Malay	111M	42M	16,696,882	6,045,753
Assamese	113M	71M	6,956,663	4,366,570	Malayalam	4.9G	2.5G	189,534,472	95,892,551
Asturian	2.4M	2.0M	381,005	325,237	Maltese	24M	17M	2,995,654	2,163,358
Avaric	409K	324K	24,720	19,478	Marathi	2.7G	1.4G	162,609,404	82,130,803
Azerbaijani	2.8G	1.5G	322,641,710	167,742,296	Galician	620M	384M	102,011,291	63,600,602
Bashkir	128M	90M	9,796,764	6,922,589	Georgian	3.6G	1.9G	171,950,621	91,569,739
Basque	848M	342M	120,456,652	45,359,710	German	308G	145G	44,878,908,446	21,529,164,172
Bavarian	503	503	399	399	Gon Konkani	2.2M	1.8M	124,277	102,306
Belarusian	1.8G	1.1G	144,579,630	83,499,037	Guarani	36K	24K	7,382	4,680
Bengali	11G	5.8G	623,575,733	363,766,143	Gujarati	1.1G	722M	72,045,701	50,023,432
Bihari	110K	34K	8,848	2,875	Haitian	3.9K	3.3K	1,014	832
Bishnupriya	4.1M	1.7M	198,286	96,940	Hebrew	20G	9.8G	2,067,753,528	1,032,018,056
Bosnian	447K	116K	106,448	20,485	Hindi	17G	8.9G	1,372,234,782	745,774,934
Breton	29M	16M	5,013,241	2,890,384	Hungarian	40G	18G	5,163,936,345	2,339,127,555
Bulgarian	32G	14G	2,947,648,106	1,268,114,977	Icelandic	1.5G	846M	219,900,094	129,818,331
Burmese	1.9G	1.1G	56,111,184	30,102,173	Ido	147K	130K	25,702	22,773
Catalan	8.0G	4.3G	1,360,212,450	729,333,440	Iloko	874K	636K	142,942	105,564
Cebuano	39M	24M	6,603,567	3,675,024	Indonesian	30G	16G	4,574,692,265	2,394,957,629
Central Bikol	885	885	312	312	Interlingua	662K	360K	180,231	100,019
Central Khmer	1.1G	581M	20,690,610	10,082,245	Interlingue	24K	1.6K	5,352	602
Central Kurdish	487M	226M	48,478,334	18,726,721	Irish	88M	60M	14,483,593	10,017,303
Chavacano	520	520	130	130	Italian	137G	69G	22,248,707,341	11,250,012,896
Chechen	8.3M	6.7M	711,051	568,146	Japanese	216G	106G	4,962,979,182	1,123,067,063
Chinese	508G	249G	14,986,424,850	6,350,215,113	Javanese	659K	583K	104,896	86,654
Chuvash	39M	26M	3,041,614	2,054,810	Kalmyk	113K	112K	10,277	10,155
Cornish	44K	14K	8,329	2,704	Kannada	1.7G	1.1G	81,186,863	49,343,462
Croatian	226M	110M	34,232,765	16,727,640	Karachay-Balkar	2.6M	2.3M	185,436	166,496
Czech	53G	24G	7,715,977,441	3,540,997,509	Kazakh	2.7G	1.5G	191,126,469	108,388,743
Danish	16G	9.5G	2,637,463,889	1,620,091,317	Kirghiz	600M	388M	44,194,823	28,982,620
Dhivehi	126M	79M	7,559,472	4,726,660	Komi	2.3M	1.2M	201,404	95,243
Dimli	146	146	19	19	Korean	24G	12G	2,368,765,142	1,120,375,149
Dutch	78G	39G	13,020,136,373	6,598,786,137	Kurdish	94M	60M	15,561,003	9,946,440
Eastern Mari	7.2M	6.0M	565,992	469,297	Lao	174M	114M	4,133,311	2,583,342
Egyptian Arabic	66M	33M	7,305,151	3,659,419	Latin	26M	8.3M	4,122,201	1,328,038
Emilian-Romagnol	25K	24K	6,376	6,121	Latvian	4.0G	1.8G	520,761,977	236,428,905
English	2.3T	1.2T	418,187,793,408	215,841,256,971	Lezghian	3.3M	3.0M	247,646	224,871
Erzya	1.4K	1.2K	90	78	Limburgan	29K	27K	4,730	4,283
Esperanto	299M	228M	48,486,161	37,324,446	Lithuanian	8.8G	3.9G	1,159,661,742	516,183,525
Estonian	4.8G	2.3G	643,163,730	309,931,463	Lojban	736K	678K	154,330	141,973
Finnish	27G	13G	3,196,666,419	1,597,855,468	Lombard	443K	433K	75,229	73,665
French	282G	138G	46,896,036,417	23,206,776,649	Low German	18M	13M	2,906,347	2,146,417
Total		6.3T	3.2T	844,315,434,723		Total	425,651,344,234		



DATA PREPARATION

Data Filtering



OSCAR 2019

Text
Deduplication

Language
Filtering

General
Cleaning

Blacklist

DATA PREPARATION

Data Filtering



OSCAR 2019

Text
Deduplication

Language
Filtering

General
Cleaning

Blacklist

- Deduplicated Oscar dataset
- Document deduplication using Locality Sensitive Hashing (LSH) with minhash

Script example for data cleaning

Locality Sensitive Hashing (LSH) with minhash

DATA PREPARATION

Data Filtering



OSCAR 2019

Text
Deduplication

Language
Filtering

General
Cleaning

Blacklist

- Language detection

DATA PREPARATION

Data Filtering



OSCAR 2019

Text
Deduplication

Language
Filtering

General
Cleaning

Blacklist

- Remove document with less than 256 characters
- Clean extra spaces and newlines

DATA PREPARATION

Data Filtering



OSCAR 2019

Text
Deduplication

Language
Filtering

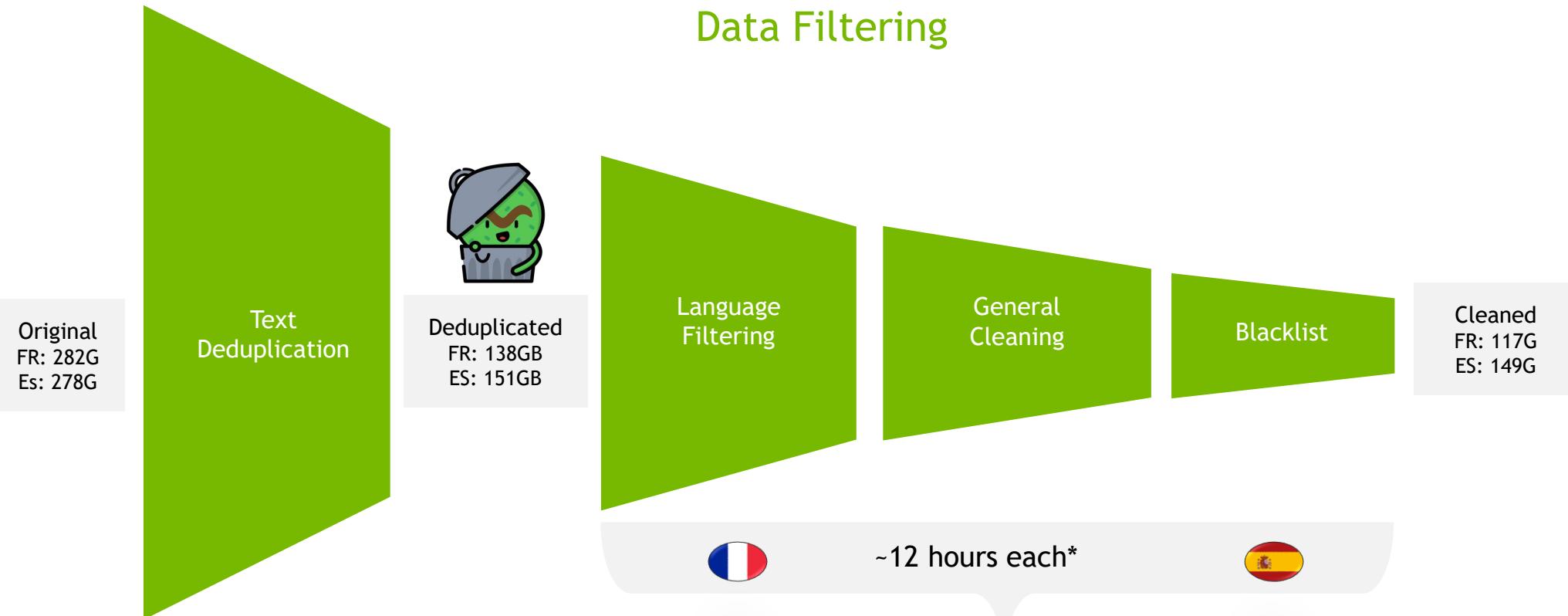
General
Cleaning

Blacklist

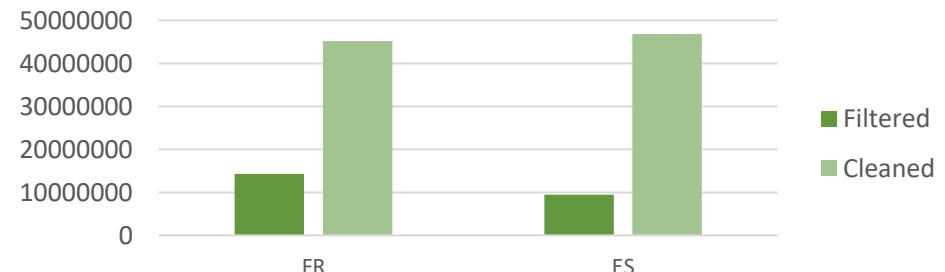
- Remove documents with at least 2 blacklisted terms
 - Fr: 592 blacklisted terms
 - Es: 214 blacklisted terms

DATA PREPARATION

Data Filtering



of filtered and cleaned documents



[Link to data cleaning script](#)



Execution time on a NVIDIA DGX-1 server:
 - dual 20-Core Intel XeonE5-2698 v4 2.2 GHz
 - 8xv100 16GBs each

DATA PREPARATION

Example of filtered documents with no text

Example of documents with no text

TRAINING THE TOKENIZER

DATA PREPARATION

Training the Tokenizer

Tokenizer



~45 min each*



```
mbendris@ukdc-dgx01:~/work/bignlp/data$
```



[Link to train GPTBPE Tokenizer example notebook](#)

Execution time on a NVIDIA DGX-1 server:
- dual 20-Core Intel XeonE5-2698 v4 2.2 GHz
- 8xv100 16GBs each

DATA CONVERSION

DATA PREPARATION

Data Conversion

Cleaned data
oscar_cleaned.jsonl

FR: 117G
ES: 149G

MMAP Format

Converted data
my-oscar_text_document.bin
my-oscar_text_document.idx

FR: 51G
ES: 71G

🇫🇷 ~30 hours* 🇪🇸 ~48 hours*

```
python Megatron-LM/tools/preprocess\_data.py \
    --input /path/to/oscar_cleaned.jsonl \
    --output-prefix my-oscar \
    --vocab /path/to/vocab.json \
    --merge-file /path/to/merges.txt \
    --tokenizer-type GPT2BPETokenizer \
    --dataset-impl mmap \
    --append-eod
```

Link to [data conversion script preprocess_data.py](#)

Execution time on a NVIDIA DGX-1 server:

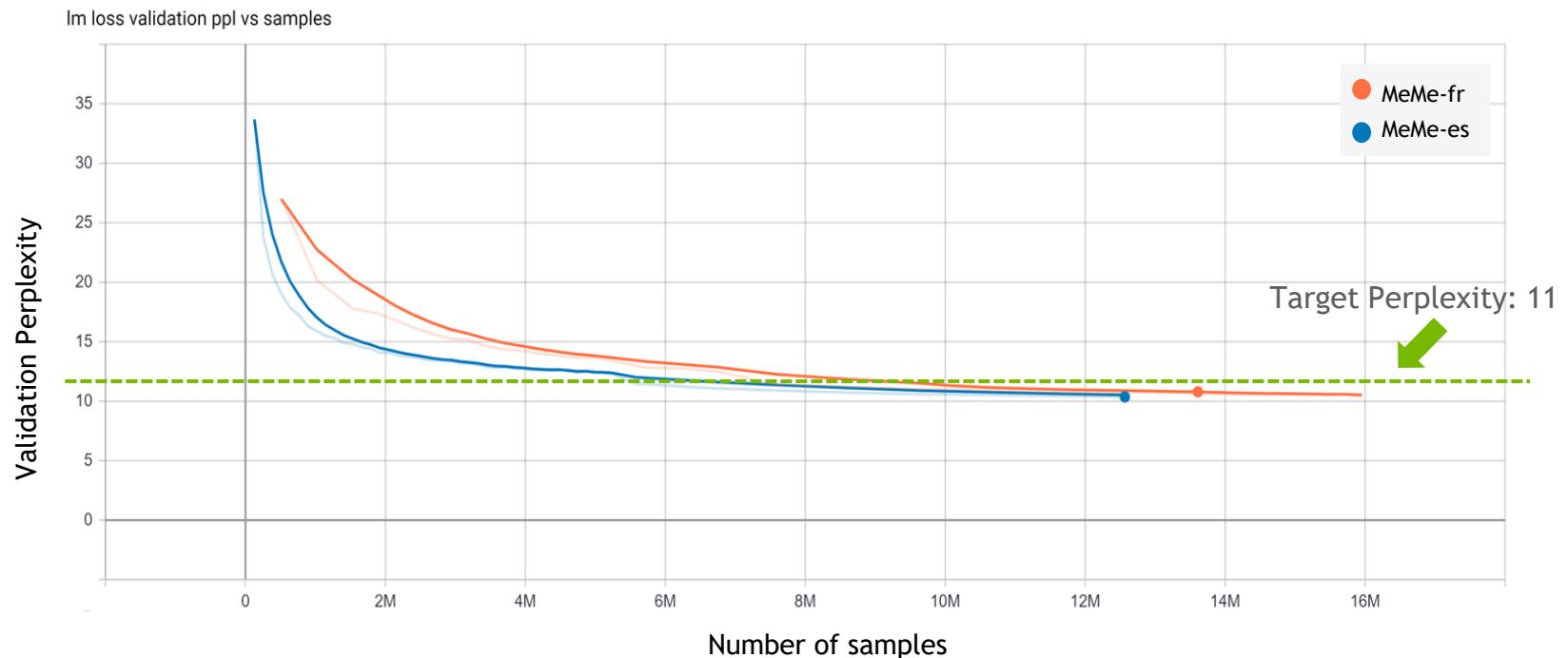
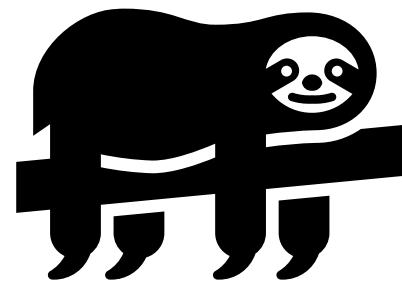
- dual 20-Core Intel XeonE5-2698 v4 2.2 GHz
- 8xV100 16GBs each

WEEK 2

TRAINING

MONITOR THE TRAINING

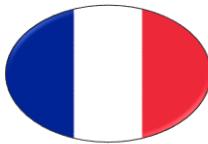
1 week of monitoring the plot below to an acceptable evaluation metric



- Code running on several DGX-1 (8 V100 32G)
- 2 to 16 nodes (subject to availability) thanks to the Megatron-LM flexibility



**~8.7 DGX A100 DAYS
~26 DGX V100 32GB (DGX-1) DAYS**



Data Preparation

Language Model Pretraining

Production Deployment

AFTER JUST 2 WEEKS OF WORK

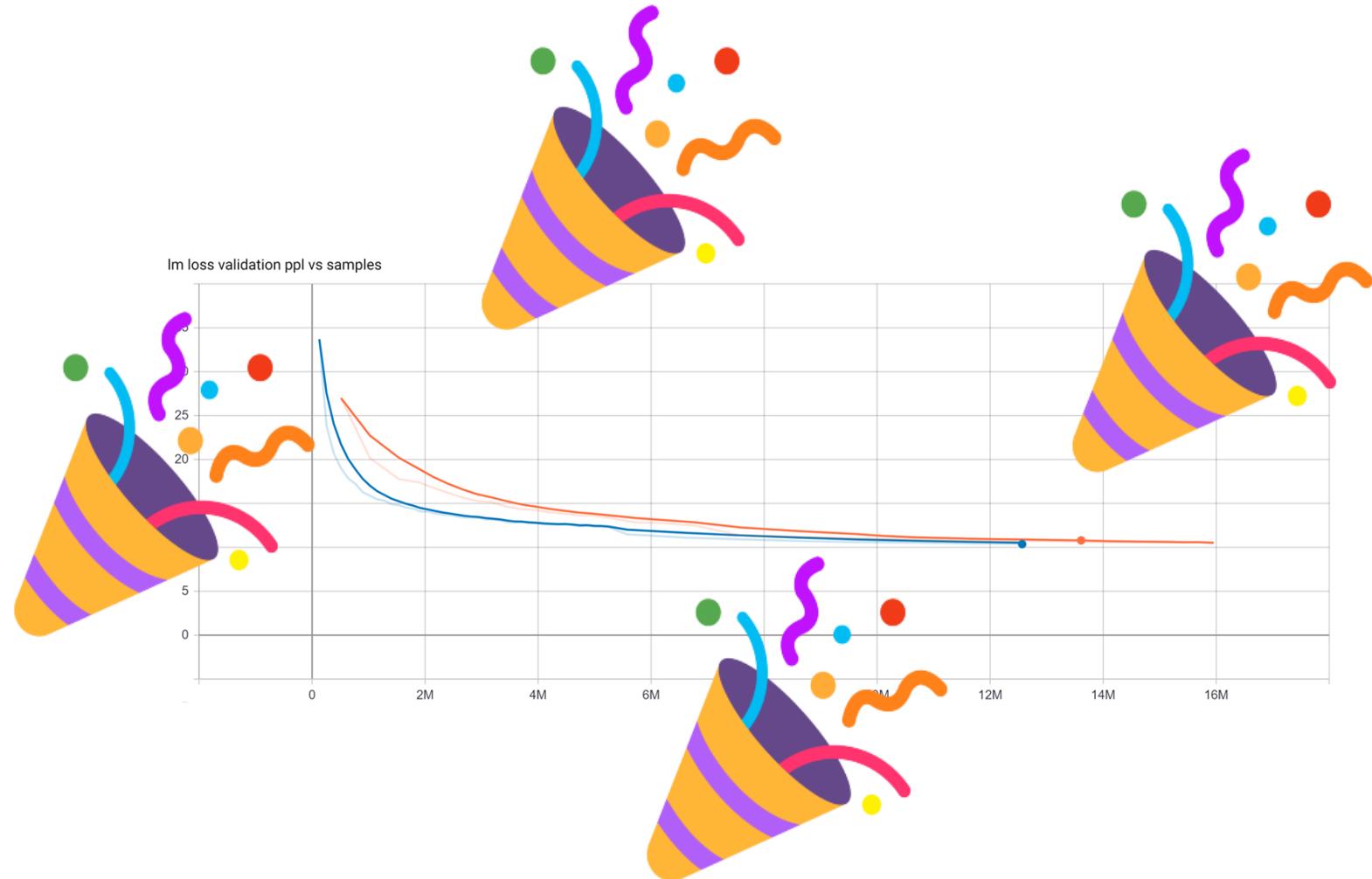
MeMe-fr / MeMe-es

MeMe-fr:

- Training LM loss: 2.410
- Validation LM loss: 2.343
- Validation PPL: 10.42

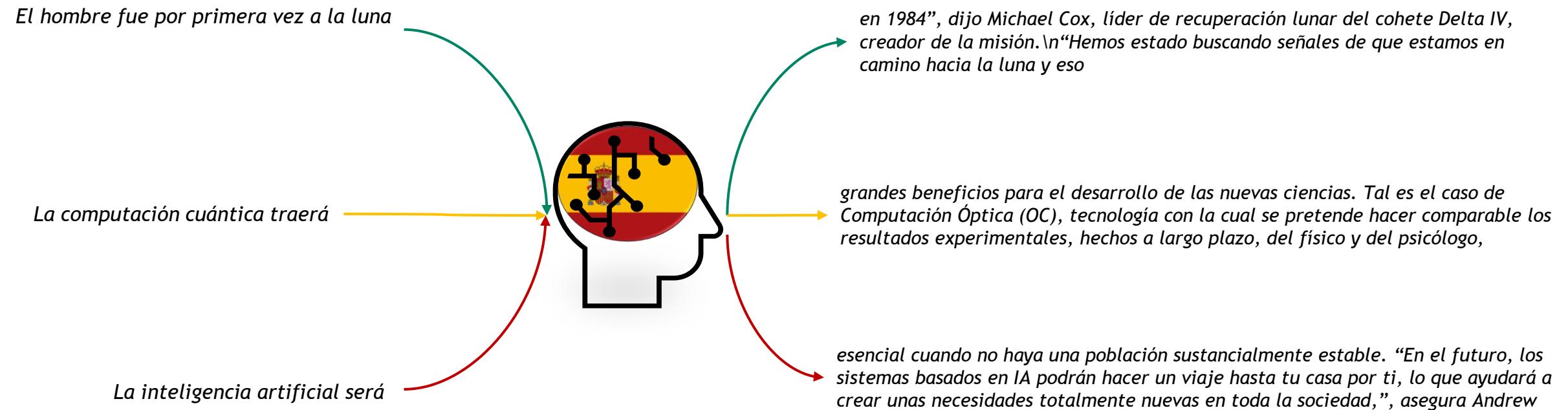
MeMe-es:

- Training LM loss: 2.404
- Validation LM loss: 2.335
- Validation PPL: 10.33



TEXT GENERATION

More Examples: MeMe-es 50 tokens



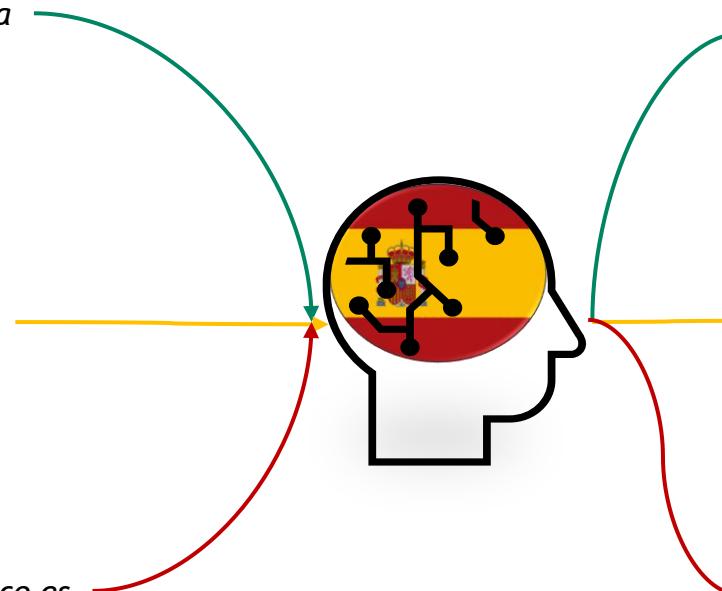
TEXT GENERATION

More Examples: MeMe-es 100 tokens

Érase una vez una tortuga

En un lugar de La Mancha,

El calentamiento climático es



de mar que vivía en un calabozo frío bajo un cielo nublado hasta que por sus exuberantes metamorfosis fue encontrada alimentándose con los desperdicios de las basuras de muchos otros animales. Esta vez nuestra tortuga no solo tuvo un mal día y atinó a comerse sus propios desechos creyendo que sería comida para ella, sino que tuvieron que hacerle cirugía reconstructiva al cerebro. ¡Estúpidos animales terriblemente tercos y desprevenidos que se miran así mismos como si

de cuyo nombre no quiero acordarme, no ha mucho tiempo que vivía un hidalgo de los de lanza en astillero, adarga antigua, rocín flaco y galgo corredor. Una olla de algo más vaca que carnero, salpicones las más noches y duelos y quebrantos los sábados lentejas; todo lo hacía el astuto de la casa. Una mañana, al levantarse, su mujer lo pilló en un arranque

atribuido a los humanos como responsable de los impactos del cambio climático tanto en las temperaturas como en los ciclos hidrológicos, ya que el globo terrestre retiene una gran superficie de agua dulce debido al calentamiento del planeta. La reducción de los tipos de lluvia está también asociada al aumento de la frecuencia de las sequías. La lluvia es un recurso importante e importante que proporciona protección contra el deshielo de los

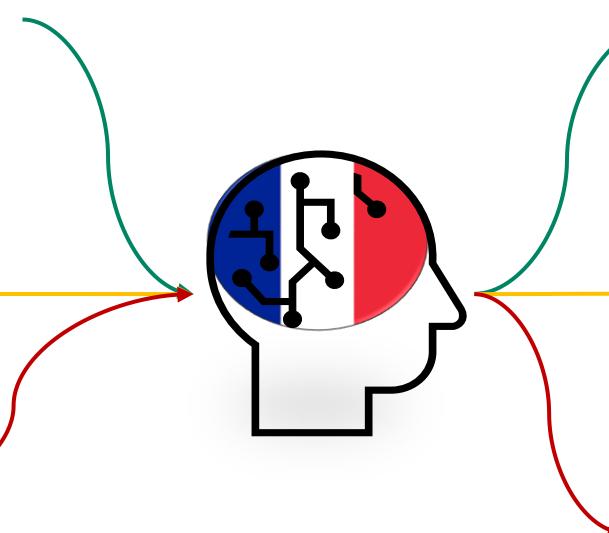
TEXT GENERATION

More Examples: MeMe-fr 50 tokens

Le président américain a rencontré le président de la république française Emmanuel

La planète Terre est

La théorie de la relativité



Macron, ce mercredi après-midi à Bruxelles. L'annonce de la visite d'Etat a été faite au terme d'une conférence de presse conjointe au site de l'aéroport à Bruxelles. Le président américain a déclaré avoir

d'ailleurs soupçonnée selon le Pentagone d'avoir très largement favorisé le système solaire à partir de 83300 ans B! Mais rien n'est prouvable visuellement. Reste que selon la NASA entre plusieurs centaines de millions et un

d'Einstein reposait déjà sur l'interaction gravitationnelle. Une modification importante de ce modèle suppose que la gravitation modifie simplement l'interaction électromagnétique et complique la représentation de la physique de gravitation, et donc celle de la gravitation quantique (théorie"

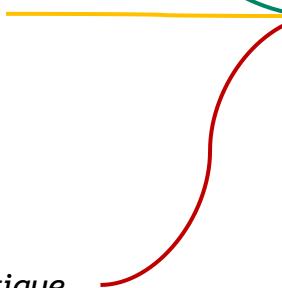
TEXT GENERATION

More Examples: MeMe-fr 100 tokens

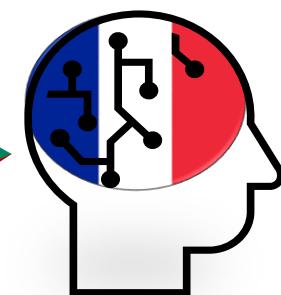
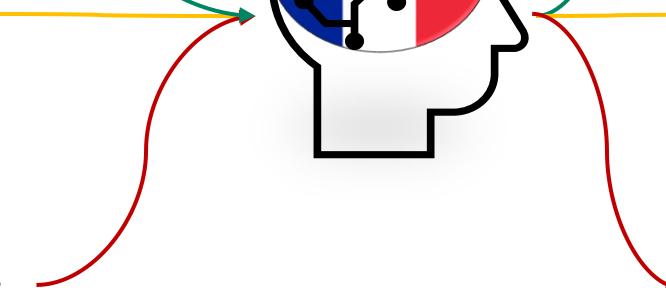
Il était une fois une tortue



Marie Curie est une physicienne



définir le réchauffement climatique



d'aquarium qui se promenait dans un parc d'attraction avec des amis... Des cosmonautes se promenaient à côté d'elle quand soudain... elle a disparu ! Apparemment une ratte farouille toute seule et après être passé entre les coussins, elle est repartie en sens inverse, courant en tous sens à toute allure !

Le troisième jour, le soleil brilla dans le ciel. La tortue rentra dans son jardin et là, elle se mit à courir,

d'excellence, disparue en 1943. Chez elle, la passion de la recherche reste l'outil indispensable de ses travaux. Directrice de l'Institut d'études nucléaires de Paris (INPL), elle est à l'origine de la mise en place des premières applications des rayons X à la radiologie médicale, notamment l'application du scanner aux radiographies du thorax. Parmi les thèses auxquelles elle a contribué, on compte, en chimie, les moyens de prévenir et de guérir par la

d'une façon simpliste et de capter ce qui aux yeux de la masse ne sert qu'à satisfaire une pulsion, n'apporte aucune donnée tangible. Aucune solution, mais au moins un espoir monumental est arrivé car la Nature n'a pas l'air si bouché depuis l'Apocalypse... Cela sonne si vrai!

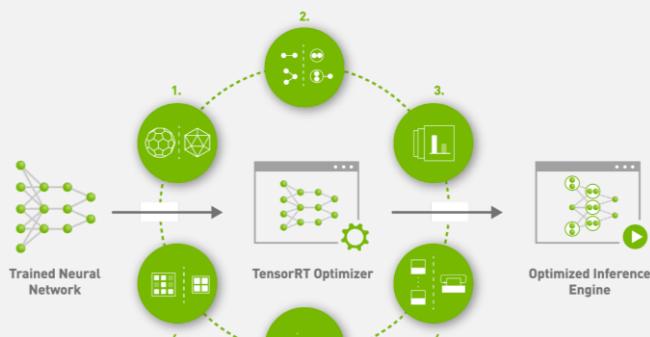
Demain qui sait si ce ne sera pas celui où nous serons menacés par le péril climatique le plus grave! Car demain, le système solaire semble avoir fermé son cœur à

WEEK 3

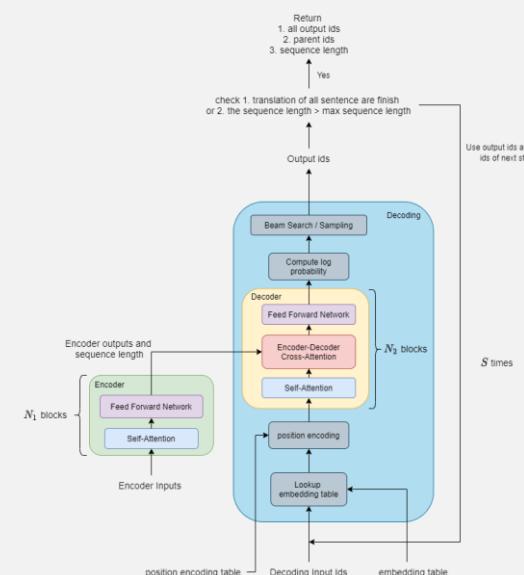
DEPLOYMENT CHALLENGES

Reduce Latency & Maximize Throughput

Model Optimization

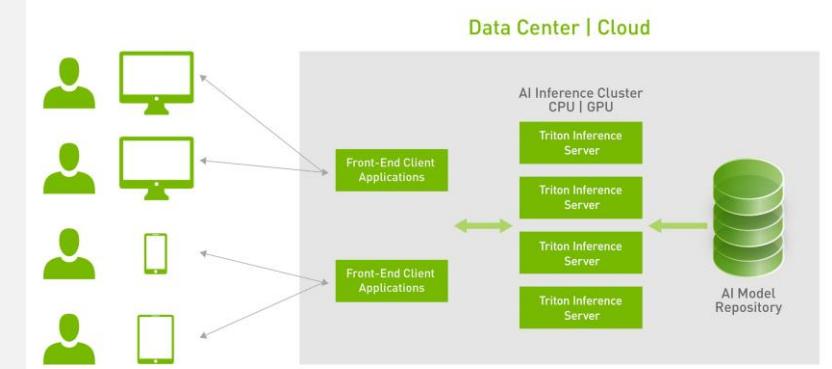


TensorRT



FasterTransformer

Model Serving



Triton Inference Server

LARGE SCALE NLP DEPLOYMENT

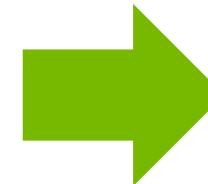
FasterTransformer and Triton Inference Server

1. Megatron to FasterTransformer Conversion

```
Python /path/to/megatron\_ckpt\_convert.py \
-i /path/to/magatron/checkpoint \
-o /path/to/converted/model/to/FT \
-t_g 2 \
-i_g 8
```

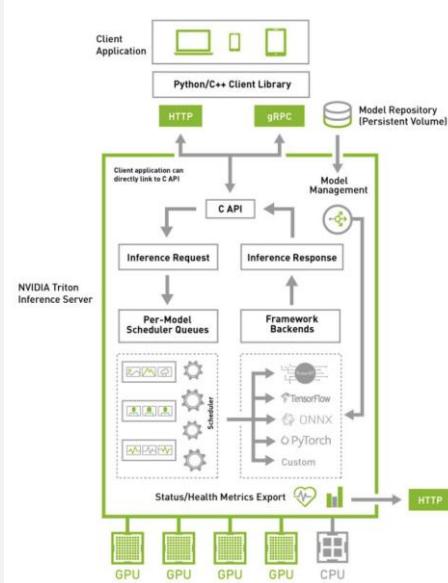
Inference with FasterTransformer

```
mpirun-n 8 --allow-run-as-root python /path/to/gpt\_sample.py \
--tensor_para_size=8 \
--layer_para_size=1 \
--ckpt_path="/path/to/converted_FT/model/to/8-gpu"
```



2. Triton Inference Server

[FasterTransformer Backend](#)



Every 2.0s: nvidia-smi

Mon Feb 21 13:02:23 2022

NVIDIA-SMI 460.32.03 Driver Version: 460.32.03 CUDA Version: 11.2						
GPU Name	Persistence-M	Bus-Id	Disp.A	Volatile Uncorr.	ECC	
Fan	Temp	Perf	Pwr/Usage/Cap	Memory-Usage	GPU-Util	Compute M. MIG M.
0 Tesla V100-SXM2...	On	00000000:06:00.0	Off	0	0%	Default N/A
N/A 34C P0 49W / 163W		7984MiB / 32510MiB				
1 Tesla V100-SXM2...	On	00000000:07:00.0	Off	0	0%	Default N/A
N/A 33C P0 44W / 163W		3MiB / 32510MiB				
2 Tesla V100-SXM2...	On	00000000:0A:00.0	Off	0	0%	Default N/A
N/A 33C P0 42W / 163W		3MiB / 32510MiB				
3 Tesla V100-SXM2...	On	00000000:0B:00.0	Off	0	0%	Default N/A
N/A 33C P0 42W / 163W		3MiB / 32510MiB				
4 Tesla V100-SXM2...	On	00000000:85:00.0	Off	0	0%	Default N/A
N/A 31C P0 42W / 163W		3MiB / 32510MiB				
5 Tesla V100-SXM2...	On	00000000:86:00.0	Off	0	0%	Default N/A
N/A 33C P0 44W / 163W		3MiB / 32510MiB				
6 Tesla V100-SXM2...	On	00000000:89:00.0	Off	0	0%	Default N/A
N/A 34C P0 44W / 163W		3MiB / 32510MiB				
7 Tesla V100-SXM2...	On	00000000:8A:00.0	Off	0	0%	Default N/A
N/A 33C P0 44W / 163W		3MiB / 32510MiB				

Processes:

GPU ID	GI ID	CI	PID	Type	Process name	GPU Memory Usage
0 N/A	N/A		2528496	C	python	7981MiB

rno1-m03-g05-dgx1-028: Mon Feb 21 13:02:23 20:~

```
root@rno1-m03-g05-dgx1-028:/megatron_workspace/bignlp-scripts/FasterTransformer/build# root@rno1-m03-g05-dgx1-028:/megatron_workspace/bignlp-scripts/FasterTransformer/build# mpirun -n 1 --allow-run-as-root python ./examples/pytorch/gpt/gpt_sample.py --tensor_para_size=1 --ckpt_path /data/checkpoints_converted/c-model_v4/iter_0053531/1-gpu/ --vocab_size 51200 --head_num 32 --layer_num 24 --vocab_file /data/vocab.json --merges_file /data/merges.txt --max_seq_len 2048 --size_per_head 72 --start_id 221 --end_id 0 --max_batch_size 1 --output_len 1 --sample_input_file /data/test_input.txt --sample_output_file /data/test_output.txt --time
=====
Arguments =====
layer_num: 24
output_len: 1
head_num: 32
size_per_head: 72
vocab_size: 51200
beam_width: 1
top_k: 1
top_p: 0.0
temperature: 1.0
len_penalty: 1.0
beam_search_diversity_rate: 0.0
tensor_para_size: 1
pipeline_para_size: 1
ckpt_path: /data/checkpoints_converted/c-model_v4/iter_0053531/1-gpu/
lib_path: ./lib/libth_parallel_gpt.so
vocab_file: /data/vocab.json
merges_file: /data/merges.txt
start_id: 221
end_id: 0
max_batch_size: 1
repetition_penalty: 1.0
max_seq_len: 2048
fp16: False
time: True
sample_input_file: /data/test_input.txt
sample_output_file: /data/test_output.txt
is_fix_random_seed: True
int8_mode: 0
```

Device Tesla V100-SXM2-32GB-LS

DONE

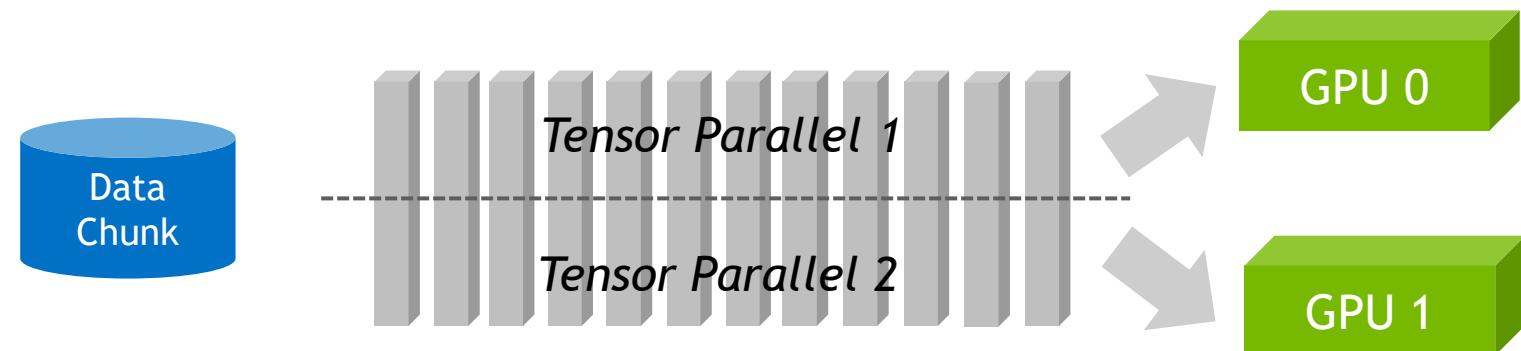
LANGUAGE MODEL PRETRAINING

DISTRIBUTED TRAINING

Architecture and Hyperparameters

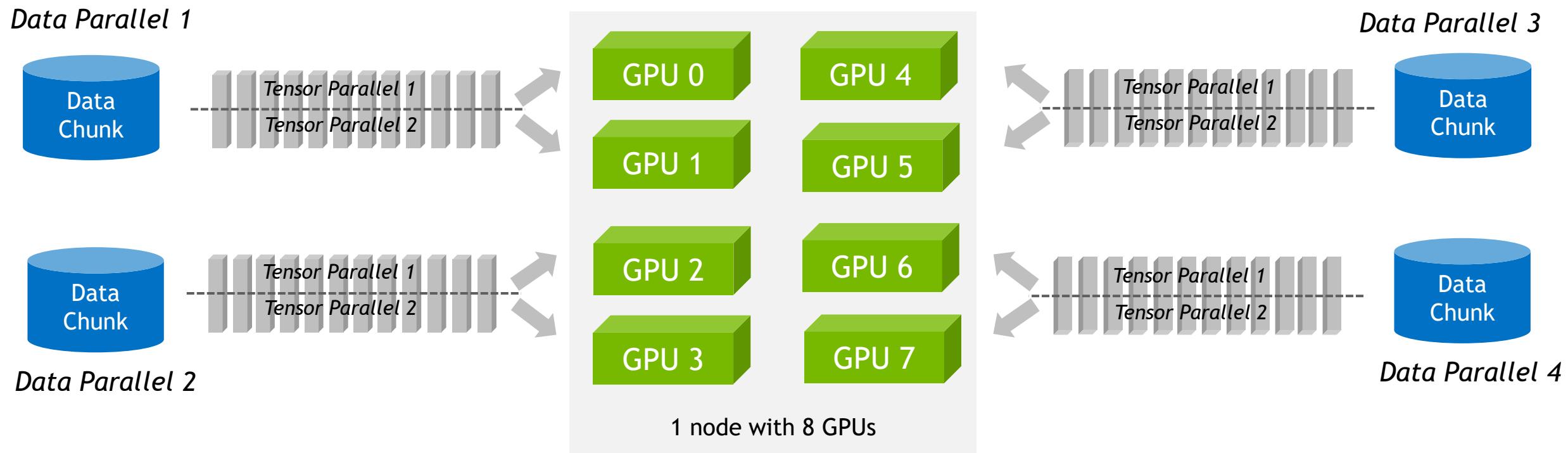
# Parameters	# Layers	# Hidden Size	Sequence Length	Vocabulary Size
1.7B	24	2304	2048	50200

LR	lr-decay-style	Weight Decay	Init method std	Micro Batch Size	Warmup Samples	Tensor Parallel	Pipeline Parallel
1.2e-4	cosine	0.1	0.014	8	30516	2	1



DISTRIBUTED TRAINING

Model Distribution within 1 node: Tensor Parallel = 2, Data parallel = 4



DISTRIBUTED TRAINING

Training the Language Model with Megatron-LM

```
GPT_ARGS=" \
  --num-layers 24 \
  --hidden-size 2304 \
  --num-attention-heads 32 \
  --seq-length 2048 \
  --max-position-embeddings 2048 \
  \
  --tensor-model-parallel-size 2 \
  --pipeline-model-parallel-size 1 \
  --exit-duration-in-mins 470 \
  --micro-batch-size 8 \
  --global-batch-size 512 \
  \
  --lr 1.2e-4 \
  --lr-decay-samples 20800000 \
  --lr-decay-style cosine \
  --lr-warmup-samples 30516 \
  --min-lr 1.2e-5 \
  --weight-decay 0.1 \
  --clip-grad 1.0 \
  --adam-beta1 0.9 \
  --adam-beta2 0.95 \
  \
  --vocab-file /data/vocab.json \
  --merge-file /data/merges.txt \
  \
  --fp16 \
  ...
"
```

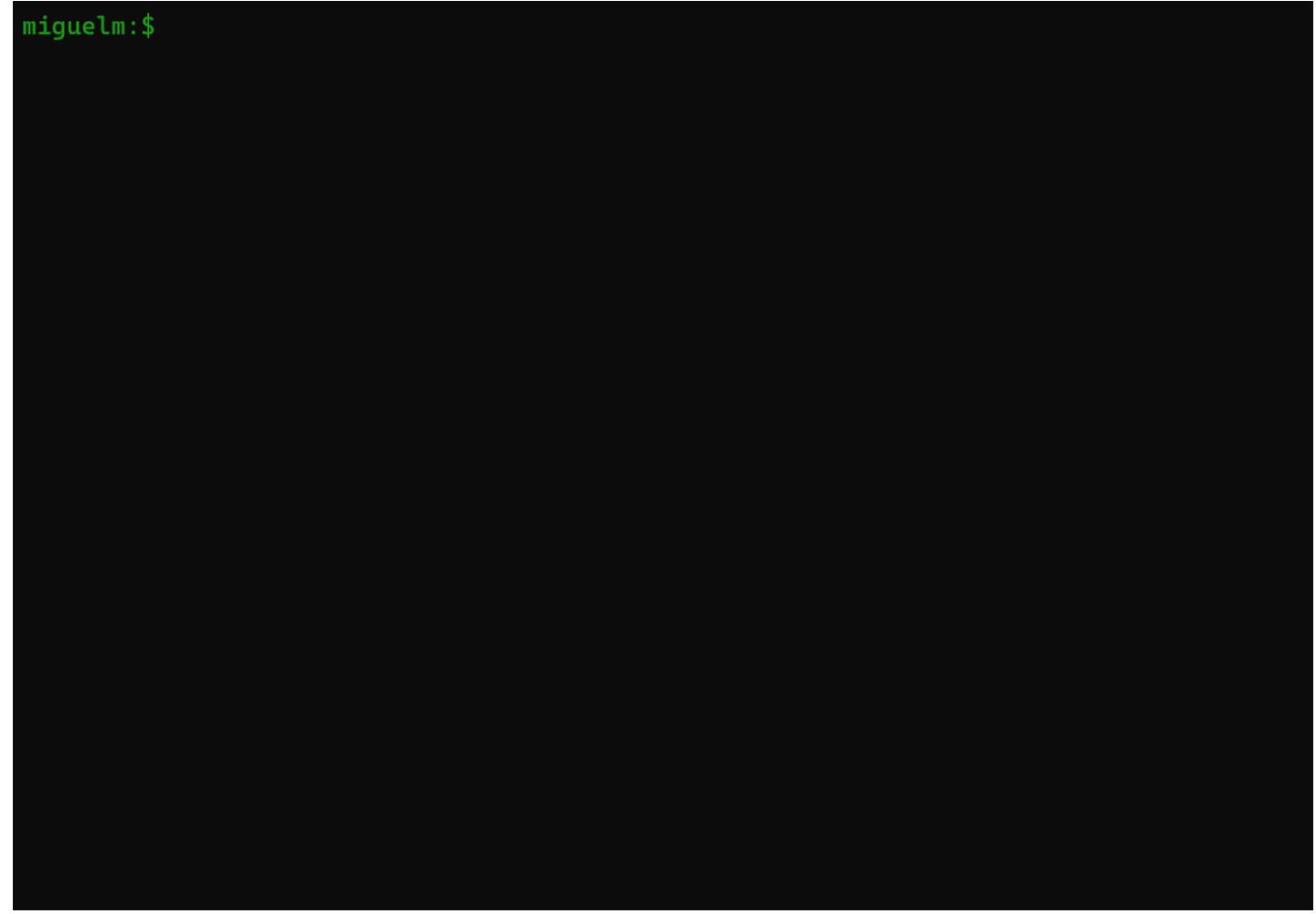
GPT Model Architecture

Strategy for Distributed Training Parameters

Optimizer Parameters

Tokenizer

FP16 Training



TEXT GENERATION

Examples: MeMe-fr



```
rbendris@ukdc-dgx01:~$ curl 'http://localhost:5000/api' -X 'PUT' -H 'Content-Type: application/json; charset=UTF-8' -d '{"prompts":["Il était une fois une tortue qui vivait dans "], "tokens_to_generate":50}' | jq -r '.text'
```



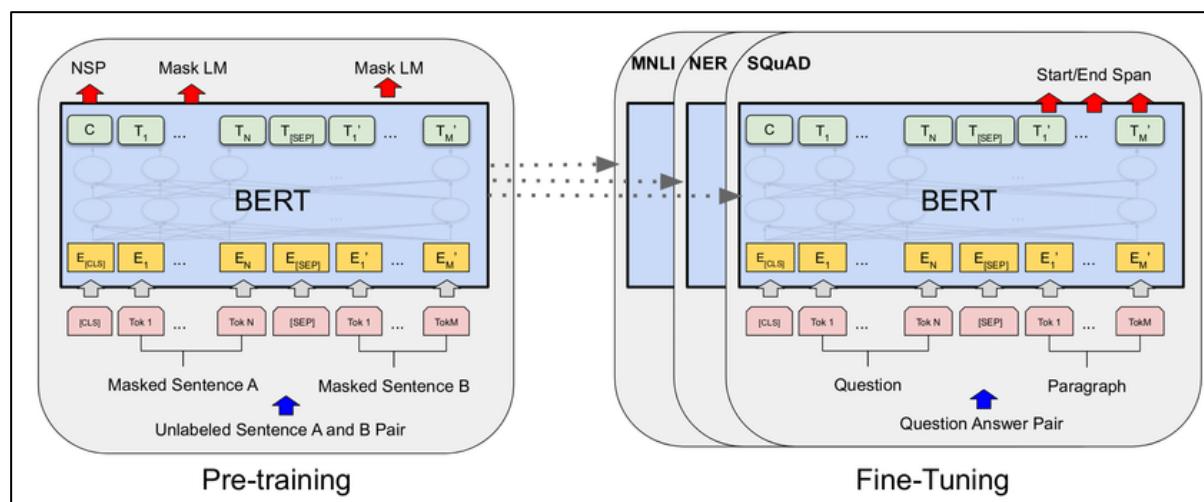
Example of REST server for GPT Text Generation

DOWNSTREAM TASKS



DOWNSTREAM TASKS

Supervised Finetuning



Traditional fine-tuning (not used for GPT-3)

Fine-tuning

The model is trained via repeated gradient updates using a large corpus of example tasks.



DOWNSTREAM TASKS

Zero/Few Shot Learners

Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

- 1 Translate English to French: ← task description
- 2 cheese => ← prompt

One-shot

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

- 1 Translate English to French: ← task description
- 2 sea otter => loutre de mer ← example
- 3 cheese => ← prompt

Few-shot

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

- 1 Translate English to French: ← task description
- 2 sea otter => loutre de mer ← examples
- 3 peppermint => menthe poivrée
- 4 plush girafe => girafe peluche
- 5 cheese => ← prompt

PROMPT ENGINEERING

How to formulate the questions?

```

mbendris@ukdc-dgx01:~$ curl 'http://localhost:5000/api' -X 'PUT' -H 'Content-Type: application/json; charset=UTF-8' -d '{"prompts":["Marie Curie est prix Nobel de"], "tokens_to_generate":1}' | jq -r '.text'
% Total    % Received % Xferd  Average Speed   Time   Time Current
          Dload  Upload Total Spent   Left Speed
100  211  100  142  100    69  3463  1682 --::--- --::--- 5146
[
  "Marie Curie est prix Nobel de physique"
]
mbendris@ukdc-dgx01:~$ curl 'http://localhost:5000/api' -X 'PUT' -H 'Content-Type: application/json; charset=UTF-8' -d '{"prompts":["Marie Curie est prix Nobel de"], "tokens_to_generate":1}' | jq -r '.text'
% Total    % Received % Xferd  Average Speed   Time   Time Current
          Dload  Upload Total Spent   Left Speed
100  207  100  138  100    69  3729  1864 --::--- --::--- 5594
[
  "Marie Curie est prix Nobel de Chimie"
]
mbendris@ukdc-dgx01:~$ curl 'http://localhost:5000/api' -X 'PUT' -H 'Content-Type: application/json; charset=UTF-8' -d '{"prompts":["Le nombre de prix Nodel pour Marie Curie est de:"], "tokens_to_generate":1}' | jq -r '.text'
% Total    % Received % Xferd  Average Speed   Time   Time Current
          Dload  Upload Total Spent   Left Speed
100  280  100  191  100    89  4547  2119 --::--- --::--- 6829
[
  "Le nombre de prix Nodel pour Marie Curie est de : 11"
]
```

DOWNSTREAM TASKS

Prompt Engineering

Type	Task	Input ([X])	Template	Answer([Y])
Text CLS	Sentiment	I love this movie.	[X] The movie is [Y]	great fantastic ...
	Topics	He prompted the LM.	[X] The text is about [Y]	sports science ...
	Intention	What is taxi fare to Denver?	[X] The question is about [Y]	quantity city ...
Text-span CLS	Aspect Sentiment	Poor service but good food.	[X] What about service? [Y]	Bad Terrible ...
Text-pair CLS	NLI	[X1]: An old man with ... [X2]: A man walks ...	Hypothesis: [X1], Premise: [X2], Answer: [Y]	Contradiction Entailment ...
Tagging	NER	[X1]: Mike went to Paris. [X2]: Paris	[X1] [X2] is a [Y]	Yes No ...
Text Generation	Summarization	Las Vegas police ...	[X] TL;DR: [Y]	The victim ... A woman
	Translation	Je vous aime.	French [X] English: [Y]	I love you. I fancy you. ...

Prompts			
manual	<i>DirectX is developed by</i> <i>y_{man}</i>	<i>y_{mine}</i> <i>released the DirectX</i>	
mined		<i>y_{mine}</i>	<i>DirectX is created by</i> <i>y_{para}</i>
paraphrased			
Top 5 predictions and log probabilities			
	<i>y_{man}</i>	<i>y_{mine}</i>	<i>y_{para}</i>
1	<u>Intel</u>	-1.06	<u>Microsoft</u> -1.77 <u>Microsoft</u> -2.23
2	<u>Microsoft</u>	-2.21	They -2.43 Intel -2.30
3	IBM	-2.76	It -2.80 default -2.96
4	Google	-3.40	Sega -3.01 Apple -3.44
5	Nokia	-3.58	Sony -3.19 Google -3.45

Figure 1: Top-5 predictions and their log probabilities using different prompts (manual, mined, and paraphrased) to query BERT. Correct answer is underlined.

ID	Modifications	Acc. Gain
P413	<i>x plays in</i> → <i>at y position</i>	+23.2
P495	<i>x was created</i> → <i>made in y</i>	+10.8
P495	<i>x was</i> → <i>is created in y</i>	+10.0
P361	<i>x is a part of y</i>	+2.7
P413	<i>x plays in</i> <i>y position</i>	+2.2

Table 6: Small modifications ([update](#), [insert](#), and [delete](#)) in paraphrase lead to large accuracy gain (%).

PRODUCTION DEPLOYMENT



PRODUCTION DEPLOYMENT

Executive Math

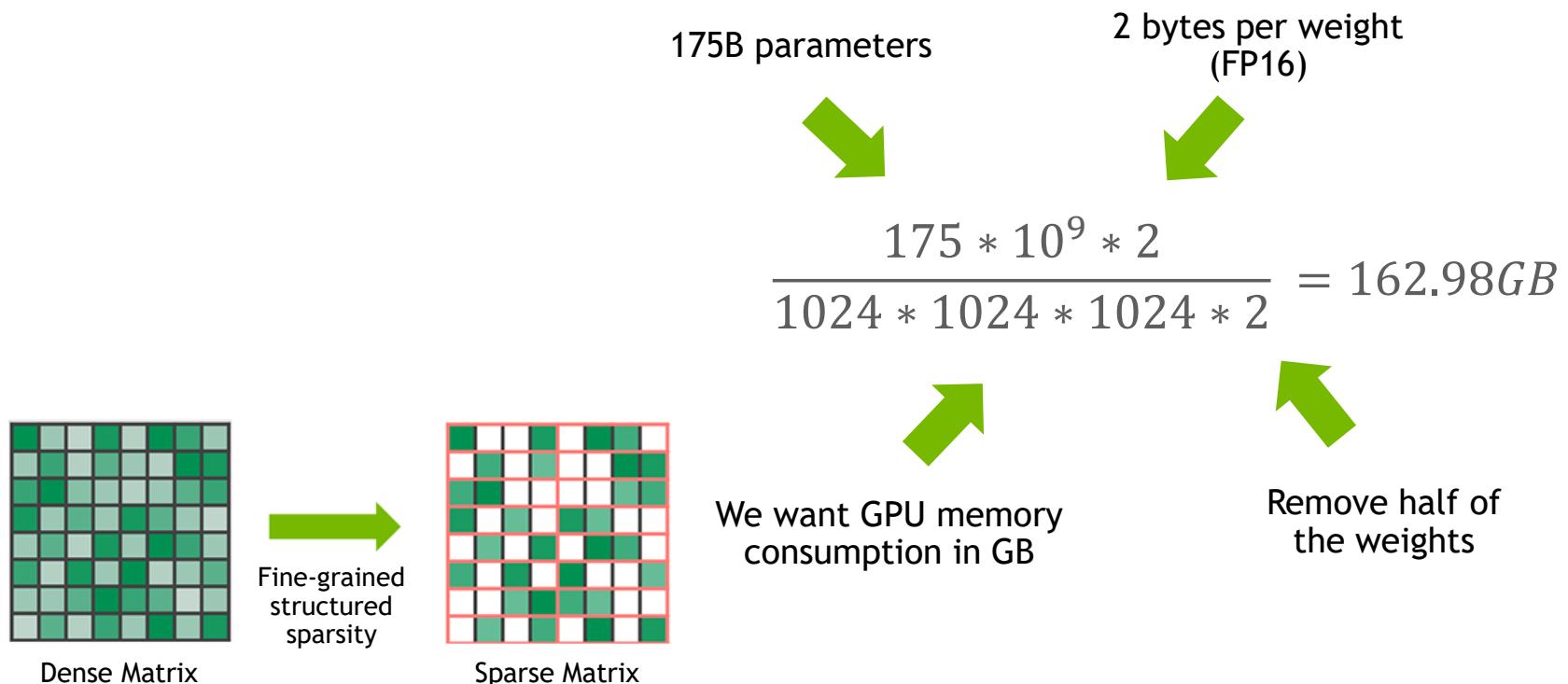
175B parameters 2 bytes per weight
(FP16)

$$\frac{175 * 10^9 * 2}{1024 * 1024 * 1024} = 325.96GB$$

We want GPU memory consumption in GB

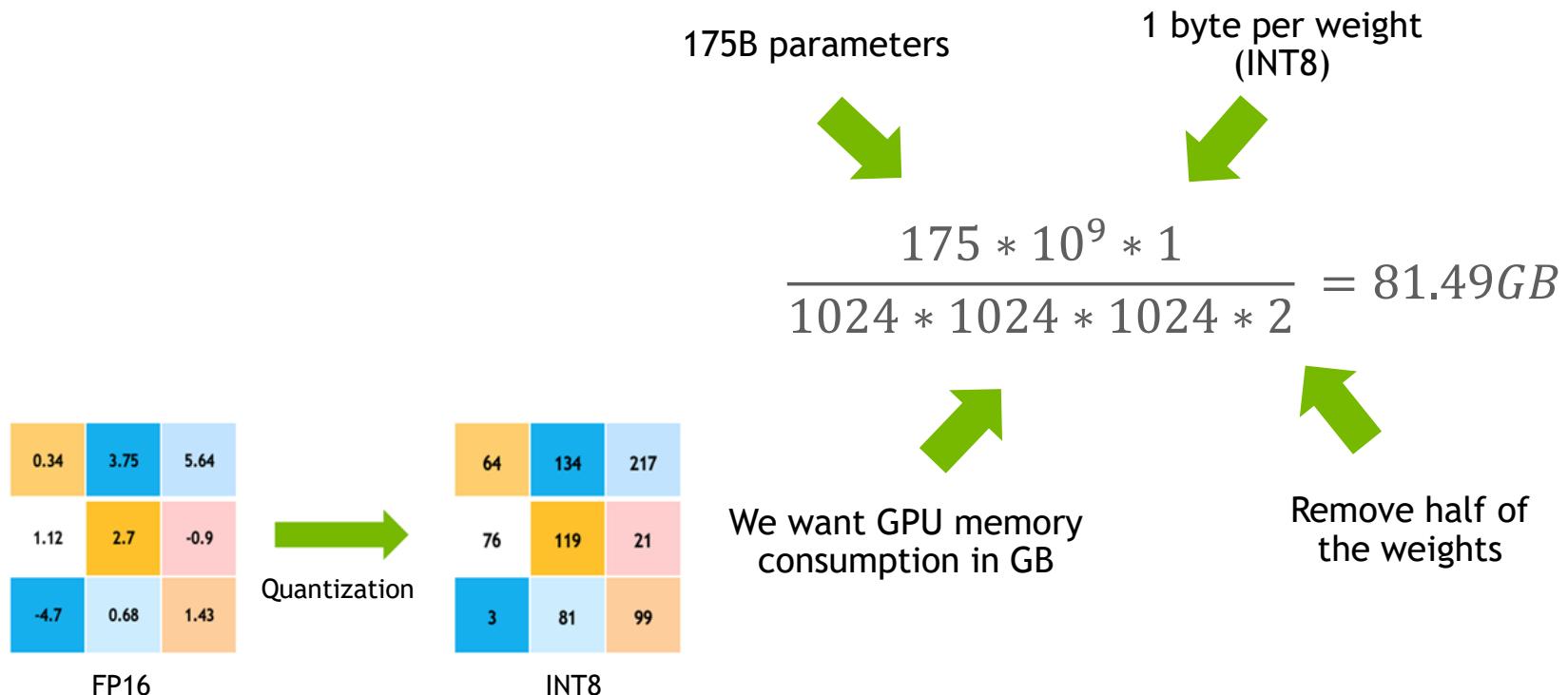
PRODUCTION DEPLOYMENT

Pruning - 2:4 Structured Sparsity



PRODUCTION DEPLOYMENT

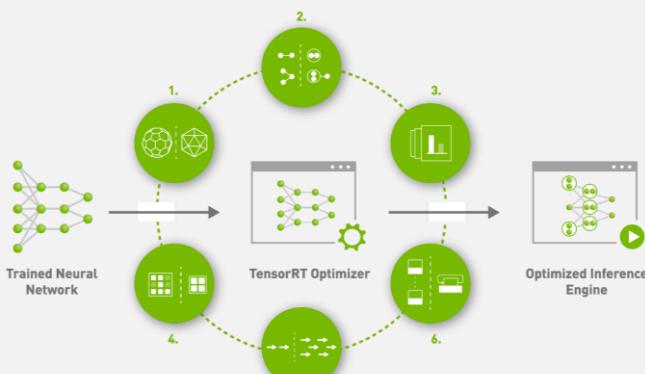
Quantization



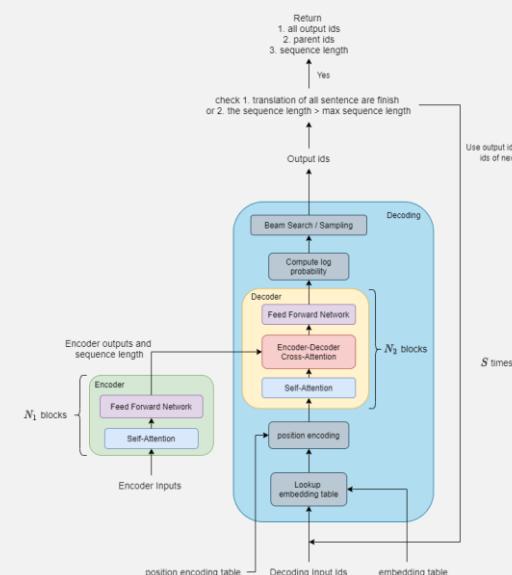
DEPLOYMENT CHALLENGES

Reduce Latency & Maximize Throughput

Model Optimization

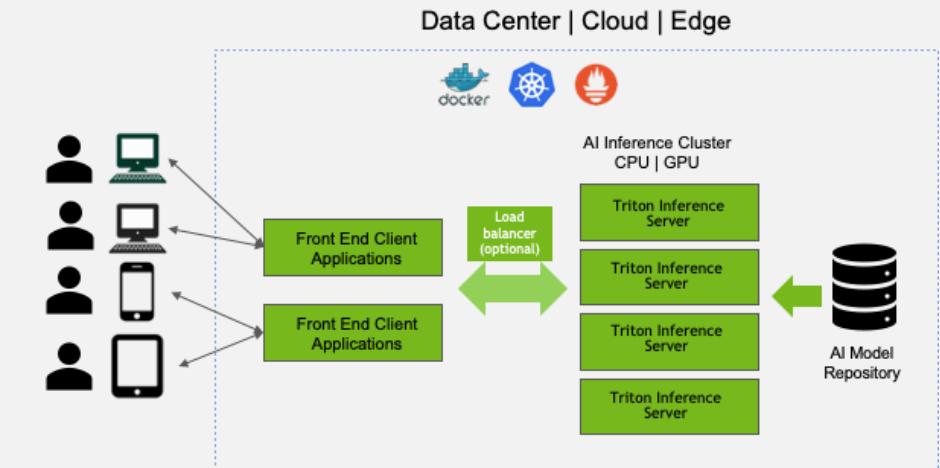


TensorRT



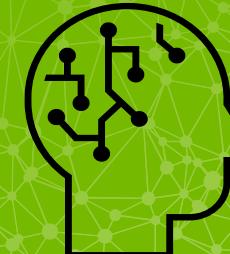
FasterTransformer

Model Serving



Triton Inference Server

UNLOCKING THE DEVELOPMENT OF LARGE-SCALE NLP MODELS



CUSTOMIZE

Scale, Improve, Downstream NLP tasks

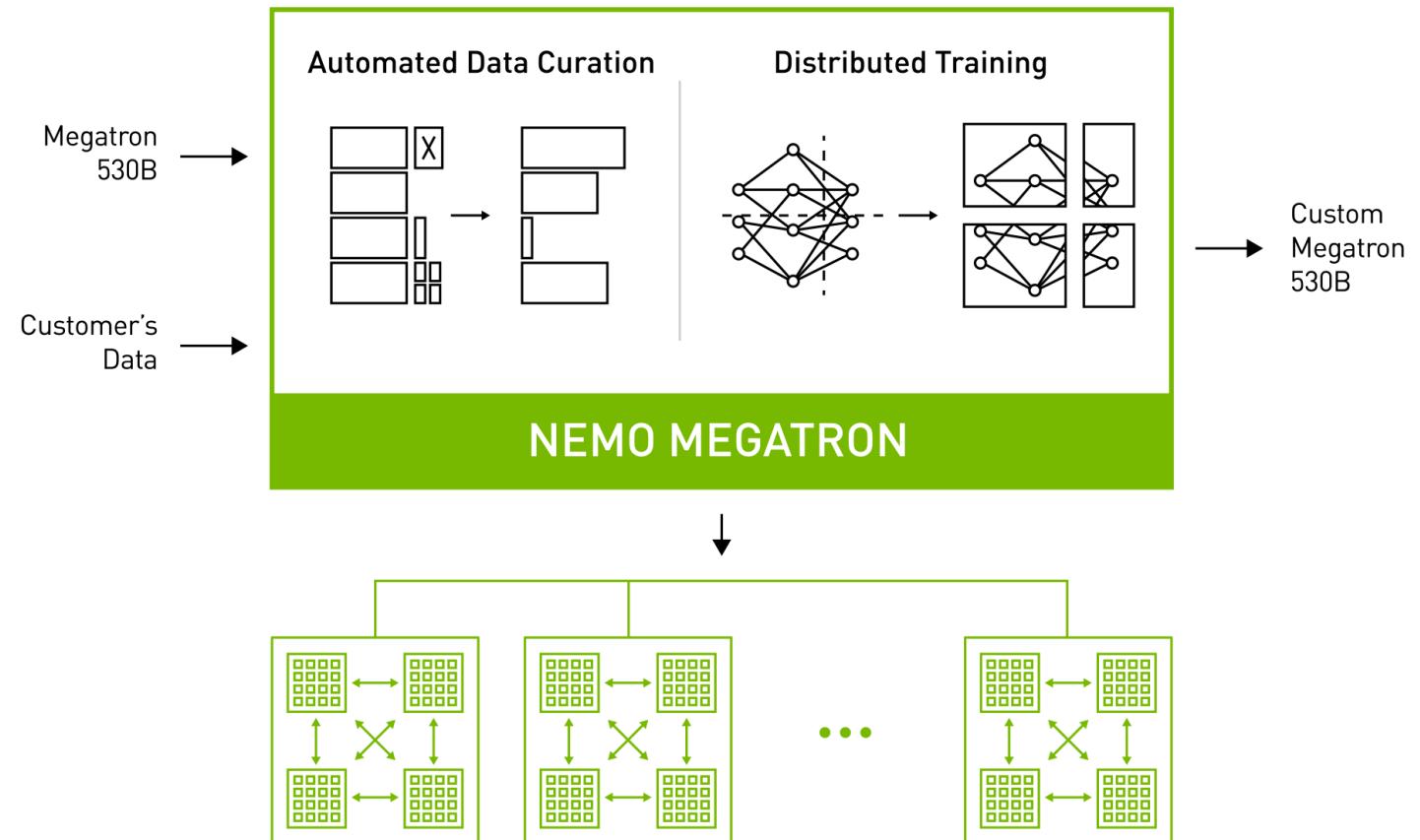


NEMO MEGATRON

Accelerated Framework for Training Large Scale NLP Models

- Scale to models with trillions of params.
- Pipeline, tensor & data parallelism.
- Automated data curation for training.
- Optimized for DGX SuperPod.
- 20B parameters model in 1 month on DGX SuperPod.

[Sign up for Early Access](#)



WRAPPING UP

OPPORTUNITIES FOR MANY LANGUAGES



GTC TALKS / PANELS

Related Live GTC22 and On-Demand Talks

[Big NLP Demystified: Business Impact of Large Language Models \[S42016\]](#) ★

In the past year we've seen unprecedented growth of models and datasets that deal with natural language processing (NLP). Models such as GPT-3 or Megatron Turing are now two orders of magnitude larger than models that just recently were state of...

Adam Henryk Grzywaczewski, Senior Deep Learning Data Scientist, NVIDIA

Industry Segment: All Industries

Primary Topic: AI Strategy for Business Leaders

[ADD TO SCHEDULE](#)

Wednesday, March 23 | 11:00 AM - 11:50 AM CET

[Large Multimodal Models are Few-shot Learners: Efficient Training and Validation of Giant Multimodal World Models with Adapters \[S42156\]](#) ★

The generalizability and few-shot capabilities of large language models (LLM) like GPT-3 have opened up new possibilities for countless innovative apps. LLMs demonstrate an impressive context and language understanding that enables them to solve...

Jonas Andrulis, Founder and CEO, Aleph Alpha

Industry Segment: All Industries

Primary Topic: Conversational AI / NLP

[ADD TO SCHEDULE](#)

Thursday, March 24 | 11:00 AM - 11:25 AM CET

[Large Multimodal Models are Few-shot Learners: Efficient Training and Validation of Giant Multimodal World Models with Adapters](#)

[Big NLP Demystified: Business Impact of Large Language Models P-tuning: An Effective Prompt Engineering Method to Significantly Improve the Performance of Your Large NLP Model](#)

[Powering AI Applications with the World's Largest Language Models](#)

[Powering AI Applications with the World's Largest Language Models \[S42062\]](#) ★

Large language models, such as AI21 Labs' Jurassic-1 and OpenAI's GPT-3, are versatile tools that developers and even non-coders can leverage to build sophisticated AI applications. We'll take a deep dive into the technology behind large language...

Dan Padnos, VP of Platform, AI21 Labs

Industry Segment: All Industries

Primary Topic: Conversational AI / NLP

[ADD TO SCHEDULE](#)

Thursday, March 24 | 9:00 AM - 9:50 AM CET



[CHALLENGES OF TRAINING AND DEPLOYING MODERN NLP \[E32245\]](#)

Meriem Bendris
Solution Architect, NVIDIA



[A STEP-BY-STEP GUIDE TO BUILDING LARGE CUSTOM LANGUAGE MODELS](#)

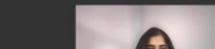
| Adam Grzywaczewski
| Meriem Bendris
| Zenodia Charpy
| Denis Timonin



ADAM GRZYWACZEWSKI
Senior Deep Learning Data Scientist, NVIDIA



YI DONG
Product Director, Hugging Face



NEHA SENGUPTA
Applied Scientist, Group42



PAVEL KALAYDIN
Director of Artificial Intelligence, Tinkoff



NVIDIA DEEP LEARNING INSTITUTE

Instructor-Led Workshops



Duration:
8 hours

Price:
N/A

Learning Objectives

Learn how to use natural language processing (NLP) Transformer-based models for "text classification" tasks, such as identifying specific types of articles from within a large library of articles or abstracts. You'll also learn how to leverage Transformer-based models for "named entity recognition (NER)" tasks, and learn how to analyze various model features, constraints, and characteristics to determine which are best suited for a particular use case based on metrics, domain specificity, and available resources.

You'll learn how to:

- Construct a Transformer neural network in PyTorch for language translation
- Build a text classification project using pre-trained BERT-variant models to classify abstracts
- Build a named-entity recognition (NER) project using pre-trained domain-specific models to identify disease names in text
- Deploy an NLP inference project to NVIDIA Triton

Upon completion, you'll be able to build NLP task applications from pre-trained language models and deploy them.

INSTRUCTOR-LED WORKSHOP

Fundamentals of Deep Learning for Multi-GPUs

[Request a workshop for your organization >](#)

[Notify me when public workshops are available >](#)



This workshop teaches you techniques for training deep neural networks on multi-GPU technology to shorten the training time required for data-intensive applications. Working with deep learning tools, frameworks, and workflows to perform neural network training, you'll learn concepts for implementing Horovod multi-GPUs to reduce the complexity of writing efficient distributed software and to maintain accuracy when training a model across many GPUs.

Learning Objectives

Workshop Details

Duration: 8 hours

Price: Contact us for pricing.

Prerequisites: Experience with gradient descent model training

BLOG POSTS

Share these with your network

DEVELOPER BLOG

DEVELOPER BLOG

<https://developer.nvidia.com/blog/applying-natural-language-processing-across-the-worlds-languages/>

<https://developer.nvidia.com/blog/using-deepspeed-and-megatron-to-train-megatron-turing-nlg-530b-the-worlds-largest-and-most-powerful-generative-language-model/>

TAKEAWAY

The slide features the NVIDIA GTC logo in the top left corner. The main title is "Build Large-Scale, Localized Language Models with NVIDIA NeMo™ Megatron". Below the title is the date "MARCH 21-24, 2022". To the right of the text are two headshots of speakers: Meriem Bendris and Miguel Angel Martinez, both from NVIDIA. They are positioned against a background of a network graph with yellow nodes and connecting lines. The text "GTC 2022 | Build Large-Scale Language Models NeMo Megatron" is at the bottom.

NVIDIA GTC

Build Large-Scale,
Localized Language
Models with NVIDIA
NeMo™ Megatron

MARCH 21-24, 2022

Meriem Bendris
NVIDIA

Miguel Angel Martinez
NVIDIA

GTC 2022 | Build Large-Scale Language Models NeMo Megatron

We are showing all the steps necessary to build and deploy 1.7B French and Spanish GPT Language Models on publicly available datasets

REACH OUT TO THE BROADER EMEA NLP TEAM

Find us on the GTC portal or LinkedIn



Meriem Bendris



Miguel Martínez



Zenodia Charpy



Denis Timonin



Adam Grzywaczewski

- ▶ Working with organisations building localized, domain specific language models
- ▶ Providing support across the end-to-end process: from scoping, to data collection, training and infrastructure deployment



*Thank
You!*