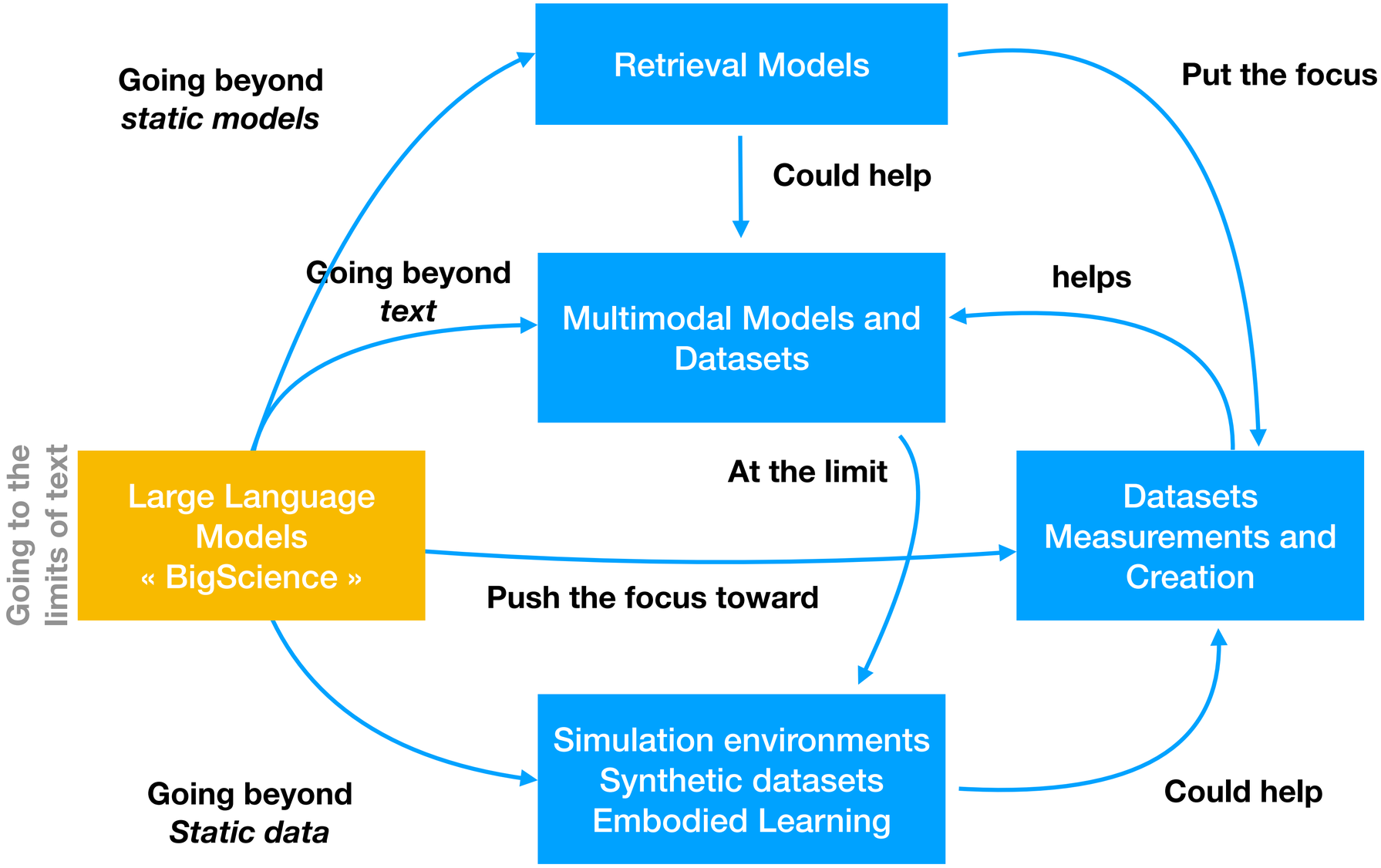
Challenges in « Natural Language Processing » research in 2022

Large models, multi-modality, retrieval and datasets

Thomas Wolf – Co-founder & CSO at Hugging Face

Hugging Face current research directions



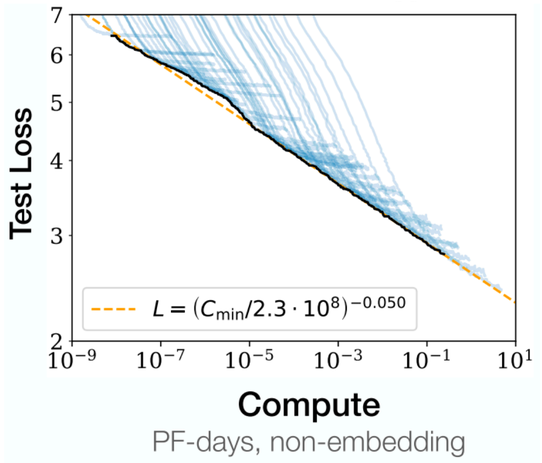
Hugging Face current research directions

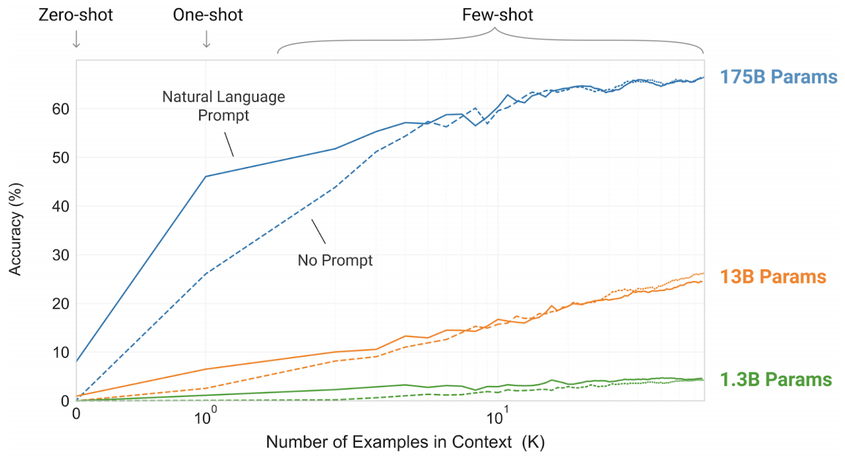
Going to the limits of text

Large Language Models « BigScience »

Large Language Models

- Following the work on scaling laws





- Several models with over 100 billion parameters trained – with SOTA and surprising results

```
To do a "farduddle" means to jump up and down really fast.  An example of a sentence that uses
the word farduddle is:
One day when I was playing tag with my little sister, she got really excited and she
started doing these crazy farduddles.
```

- Expensive to train: typically $4-8M – 4 months on 400GPUs

Lead to strong interest in big labs & startups

Large Language Models

- all 100B+ models trained up to now are **closed access**

VB VentureBeat

Naver trained a ‘GPT-3-like’ Korean language model

Naver claims the system learned 6,500 times more Korean data than OpenAI's ... Some experts believe that while HyperCLOVA, GPT-3, PanGu-α, ...

1 Jun 2021

TechCrunch

Anthropic is the new AI research outfit from OpenAI’s Dario Amodei, and it has $124M to burn

Anthropic, as it's called, was founded with his sister Daniela and its goal is to create “large-scale AI systems that are steerable, ...

28 May 2021

VB VentureBeat

AI21 Labs trains a massive language model to rival OpenAI’s GPT-3

“AI21 Labs was founded to fundamentally change and improve the way people read and write. Pushing the frontier of language-based AI requires ...

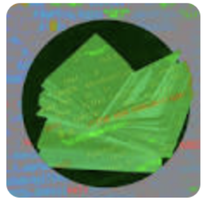
1 month ago

FC Fast Company

Ex-Googlers raise $40 million to democratize language AI

This story has been updated with more information about Cohere's approach to responsible AI. About the author. Fast Company Senior Writer Mark ...

2 days ago



Why is this a problem?

Large Language Models

➡ **Research**

- Hard to do research: *no access to data, checkpoints, internals*
- Academic researchers: *not involved*
- Lack of fields diversity: *English/Chinese, ML focused*

➡ **Environmental**

- Training similar models: *Duplication of energy*
- Carbon footprint: *Not documented*

➡ **Ethical and societal around datasets/design**

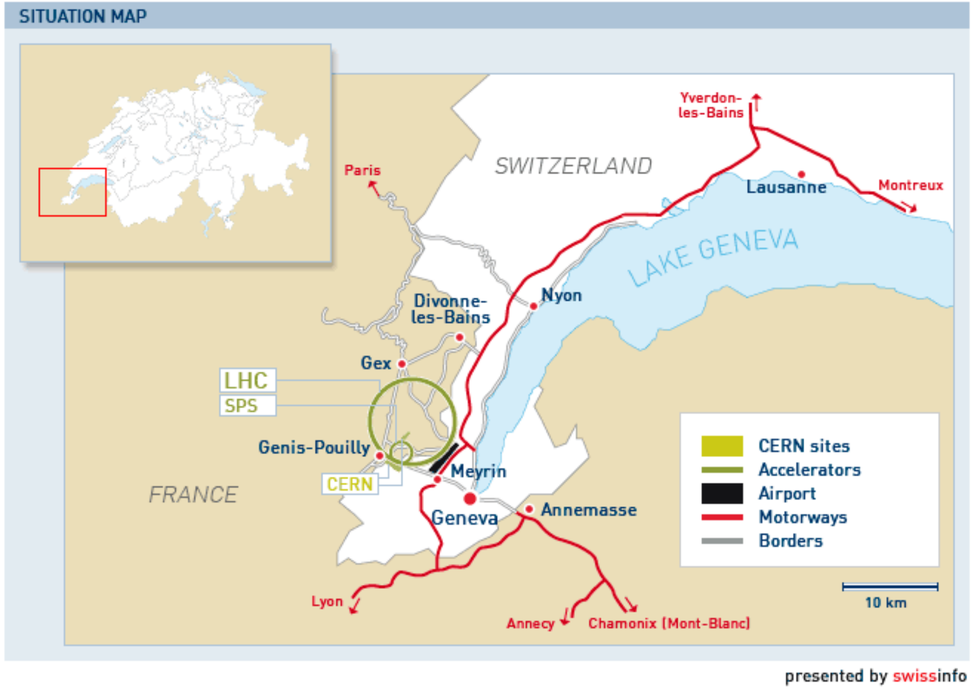
- Shortcomings in the training datasets: *Representativeness, stereotypes, PII*
- Ethical/bias/usage questions: *Only asked « a-posteriori »*

BigScience

The example of particle physics

CERN: Large Hadron Collider

- involved **10.000 researchers**
- from **100 countries**
- discovery of **59 hadrons**
- more than **2.800 papers** (😱)



World-scale research collaborations create **research tools** which are essential for science: LHC, ITER, ISS, etc

Time for similar **large, diverse, open research collaboration** in AI?

BigScience

Gather and invite a world-size research community:

- List research questions & what's needed to answer them
- Ask research questions 'a-priori' rather than 'a-posteriori'

Create and share:

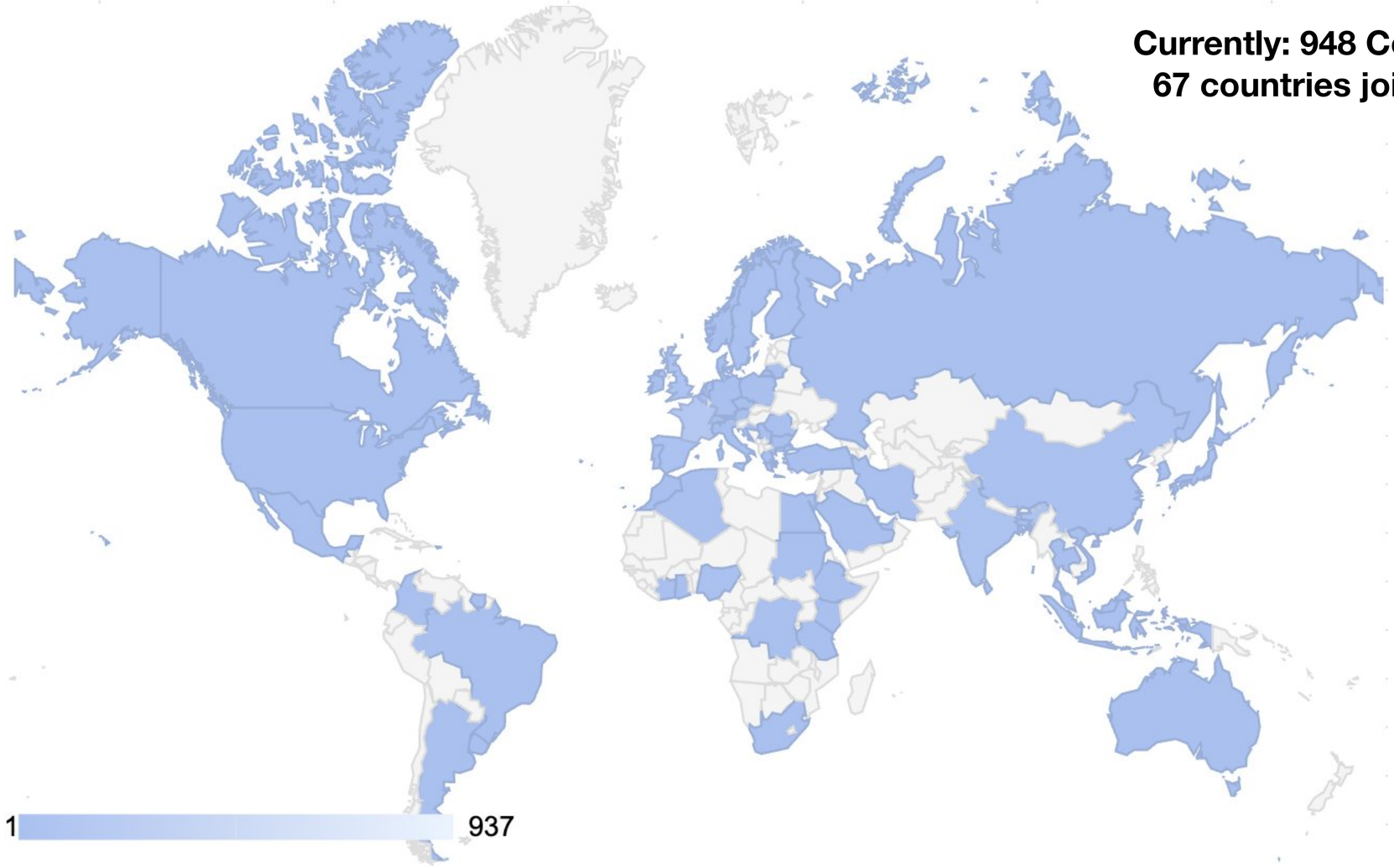
- A huge multilingual corpus
  Responsible, diverse, mindful of ethical and legal issues
- A huge multilingual language model
  Accessible to every researchers
- Code tools associated to these artifacts

Document and share all processes/discussion: *open-science*

Were people interested?

BigScience

Currently: 948 Collaborators from 67 countries joined BigScience



Co-Chair BigScience Data Sourcing and Representativeness

Pedro Ortiz Suárez

Pedro Ortiz Suárez is a PhD student at Sorbonne Université and at the ALMAnaCH research team at Inria, where he is supervised by Laurent Romary and Benoît Sagot. He is interested in large corpora and language modeling, especially for under-resourced and historical languages. He is the leader of the OSCAR project and one of the main authors of the CamemBERT language model for French.

@pjox13

Co-Chair BigScience Tokenization

Samson Tan

Samson Tan is a final year industrial PhD in computer science at the National University of Singapore and Salesforce Research Asia, advised by Min-Yen Kan and Shafiq Joty. His research lies at the intersection of language, fairness, and robustness: reducing linguistic discrimination by improving the ability of NLP systems to handle sociolinguistic variation.

@samsontmr

Co-Chair BigScience Social Impact Across Groups

Margot Mieskes

Margot Mieskes is a professor for Natural Language Processing at the University of Applied Sciences, Darmstadt. Apart from research in the area of spoken and written summarization and evaluation of summarization, she is doing a lot of work in the context of ethical issues in NLP, especially during review processes and on questions of transparency and replicability of research results.

Margot Mieskes

Co-Chair BigScience Extrinsic Evaluation

Thomas Scialom

A former banker, Thomas Scialom is now a researcher in Artificial Intelligence and a partner of reciTAL, a software editor for automatic document processing. He is the author of numerous scientific publications accepted in the most prestigious AI conferences in the world (NeurIPS, EMNLP...). Thomas also teaches at ESILV, Sorbonne University and Catalix.

@ThomasScialom

Co-Chair BigScience Sharing and Accessibility of Model and Dataset

Aaron Gokaslan

Aaron Gokaslan is currently pursuing a PhD in Computer Science at Cornell University advised by Kavita Bala. His research focuses on generative models, embodied AI, and novel view synthesis. Previously, he worked as an AI Resident at Facebook AI Research (FAIR) under Dhruv Batra. He completed his Bachelors and Masters at Brown University, advised by James Tompkins. Aaron is also the co-creator of OpenGPT2 and the OpenWebText Corpus.

@SkyLi0n

Co-Chair BigScience Engineering/Scaling

Minjia Zhang

Minjia Zhang is a Principal Researcher at Microsoft. He works on developing highly efficient systems/libraries/algorithms for accelerating large scale deep learning training and inference on parallel and distributed systems. His research results are used in multiple Microsoft systems and products, such as Bing, Ads, Windows, AzureML to improve performance and capacity

Minjia Zhang

Co-Chair BigScience Extrinsic Evaluation

Verena Rieser

Verena Rieser leads research on Conversational AI and Natural Language Generation. She is a professor at Heriot-Watt University Edinburgh, co-founder of ALANA AI and holder of a Leverhulme Senior Research Fellowship by the Royal Society. Her interests include ethical and social risks of ConvAI, continual multimodal learning and human-centred system development.

@verena_rieser

Co-Chair BigScience Evaluation / Few-Shot Generalization

Ellie Pavlick

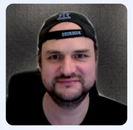
Ellie Pavlick leads the Language Understanding and Representation (LUNAR) Lab at Brown University and is a Research Scientist at Google. Her primary research interests concern natural language semantics. Her lab's current work focuses on understanding and evaluating the conceptual representations learned by neural network language models.

@Brown_NLP

Co-Chair BigScience Interpretability and Interaction/Visualization

Hendrik Strobelt

Hendrik Strobelt is a research scientist at IBM Research and explainability lead at the MIT-IBM AI Lab. His background is in visualization with a focus on interactive explainability and debugging tools for AI/ML/NLP. His work has been published at venues like IEEE VIS, ICLR, Siggraph, ACL, NeurIPS, ICCV, Nature BME, and Science Adv.



@hen_str

Co-Chair BigScience Carbon Footprint

Sasha Luccioni

Sasha is a Research Scientist at Hugging Face. Her work studies the societal and ethical impacts of AI, and her goal is to find ways to maximize the positive effects of AI while minimizing the negative ones. Sasha's work has been featured in various news and media outlets such as MIT Technology Review, WIRED and the Wall Street Journal. She is also a 2020 National Geographic Explorer and a founding member of Climate Change AI.

@SashaMTL

Co-Chair BigScience Prompt Engineering

Stephen Bach

Stephen Bach is an assistant professor of computer science at Brown University. His research focuses on weakly supervised, zero-shot, and few-shot learning. The goal of his work is to create methods and systems that drive down the labor cost of new machine learning models.

@stevebach

Co-Chair BigScience Legal and Ethical Scholarship

Carlos Muñoz Ferrandis

Carlos is a lawyer and PhD Researcher focused on the interactions between open source and standards, from an IP and competition/antitrust law angle. Although his PhD research mainly focuses on the telecoms networks industry, he is currently researching on the strategic role of open source licensing for platform leadership in AI-related markets

@Carlos_MFerr

Co-Chair BigScience Sharing and Accessibility of Model and Dataset

Danish Contractor

Danish Contractor is a Researcher at IBM Research AI in India. His research interests lie in Question-Answering & Dialog Systems, as well as, in the use of licensing mechanisms for promoting the responsible-use of AI. Danish has a PhD in Computer Science from the Indian Institute of Technology Delhi and an M.Phil from the University of Cambridge. In 2018, he was named one of the top Innovators Under 35 in India by MIT Technology Review Magazine and Mint.

@danish_c

Co-Chair BigScience Engineering/Scaling

Max Ryabinin

Max Ryabinin is a Research Scientist at Yandex and a first-year PhD student at HSE University. His work spans several topics at the intersection of NLP and ML from uncertainty estimation to the analysis of multilingual models. One of his core research interests is large-scale distributed DL, particularly on commodity devices over the Internet.

@m_ryabinin

Co-Chair BigScience Legal and Ethical Scholarship

Giada Pistilli

Giada Pistilli is a third-year PhD student in philosophy at Sorbonne Université, advised by Anouk Barberousse. Her research focuses on the ethics of Natural Language Processing, with an emphasis on empirical research on conversational agents and Large Language Models. She is also a research engineer for a chatbot company and a lecturer in several Engineering universities in France.

@GiadaPistilli

Co-Chair BigScience Modeling – Multilinguality

Hady Elsahar

Hady Elsahar is a Research Scientist at Naver Labs Europe. His main research interests are Natural Language Generation under unsupervised and low-resourced conditions, with a recent focus on Energy-Based Models, Reinforcement Learning, and Monte Carlo methods for controlling large Language Models. He holds a Ph.D. from the Université de Lyon in France and currently serves as a Board Member of the Masakhane NLP community.

@hadyelsahar

Co-Chair BigScience Intrinsic Evaluation

Sebastian Gehrmann

Sebastian is a Senior Research Scientist at Google, working on natural language generation. He currently leads the Generation Evaluation and Metrics benchmark initiative aiming to improve how NLG models are evaluated, and hopes to apply its insights to models in BigScience.

@SebGehr

Co-Chair BigScience Data Sourcing and Representativeness

Zeerak Talat

Zeerak is a Post-doc at The Digital Democracies Institute, SFU and is the co-chair with Angelina McMillan-Major and Pedro Ortiz Suarez for the data sourcing working group in the BigScience initiative. Zeerak's research focuses on the foundational limitations and the ethics of machine learning and NLP technologies as viewed through content moderation and social prediction tasks.

@Zeerak Talat

Co-Chair BigScience Language Models for Historical Texts

Daniel van Strien

Daniel van Strien is a Digital Curator at the British Library working on the Living with Machines project. Daniel is particularly interested in how to make sure GLAM institutions (Galleries, Libraries, Archives and Museums) can benefit from and support machine learning, by making machine learning methods more accessible and valuable for domain experts.

@vanstriendaniel

Co-Chair BigScience Modeling – Architecture and Scaling

Iz Beltagy

Iz Beltagy is a senior research scientist at AI2. His research focuses on language models pretraining, evaluation, and efficiency, long document processing, domain adaptation, and summarization. His recent research received best paper award at AKBC2021 and best paper runner-up at ACL2020.

@i_beltagy

59 chairs and 30 Working Groups

And so many more…

BigScience

Jean Zay public (gov funded) supercomputer at IDRIS (South of Paris, France)

Accelerated partition (or GPU partition)

- 261 four-GPU accelerated compute nodes with:
  - 2 Intel Cascade Lake 6248 processors (20 cores at 2.5 GHz), namely 40 cores per node
  - 192 GB of memory per node
  - 4 Nvidia Tesla V100 SXM2 GPUs (32 GB)
- 31 eight-GPU accelerated compute nodes, currently dedicated to the AI community with:
  - 2 Intel Cascade Lake 6226 processors (12 cores at 2.7 GHz), namely 24 cores per node
  - 20 nodes with 384 GB of memory and 11 nodes with 768 GB of memory
  - 8 Nvidia Tesla V100 SXM2 GPUs (32 GB)
- Extension in the summer of 2020, 351 four-GPU accelerated compute nodes with:
  - 2 Intel Cascade Lake 6248 processors (20 cores at 2.5 GHz), namely 40 cores per node
  - 192 GB of memory per node
  - 4 Nvidia Tesla V100 SXM2 GPUs (16 GB)

Jean Zay 3

Given the size of the BigScience project and its predicted wide impact, strong support could be found to increase the size of the public cluster itself with the addition of 52 HPE Apollo 6500 Gen 10 servers in December 2021 with the following configuration:

- GPUs: **416 A100 80GB** GPUs (52 nodes) - using 384 gpus (48 nodes) and keeping 32 gpus (4 nodes) in reserve
- 8 GPUs per node Using NVLink 4 inter-gpu connects, 4 OmniPath links
- CPU: AMD
- CPU memory: 512GB per node
- GPU memory: 640GB per node
- Inter-node connect: Omni-Path Architecture (OPA)
- NCCL-communications network: a fully dedicated subnet
- Disc IO network: shared network with other types of nodes

BigScience

- Outcomes

Between words and characters:
A Brief History of Open-Vocabulary Modeling and Tokenization in NLP

Sabrina J. Mielke [1,2] **Zaid Alyafeai** [3] **Elizabeth Salesky** [1]
Colin Raffel [2] **Manan Dey** [4] **Matthias Gallé** [5] **Arun Raja** [6]
Chenglei Si [7] **Wilson Y. Lee** [8] **Benoît Sagot** [9*] **Samson Tan** [10*]
BigScience Workshop Tokenization Working Group

[1]Johns Hopkins University [2]HuggingFace [3]King Fahd University of Petroleum and Minerals [4]SAP
[5]Naver Labs Europe [6]Institute for Infocomm Research, A*STAR Singapore [7]University of Maryland
[8]BigScience Workshop [9]Inria Paris [10]Salesforce Research Asia & National University of Singapore
sjm@sjmielke.com

Abstract

What are the units of text that we want to model? From bytes to multi-word expressions, text can be analyzed and generated at many granularities. Until recently, most natural language processing (NLP) models operated over words, treating those as discrete and atomic tokens, but starting with byte-pair encoding (BPE), subword-based approaches have become dominant in many areas, enabling small vocabularies while still allowing for fast inference. Is the end of the road character-level model or byte-level processing? In this survey, we connect several

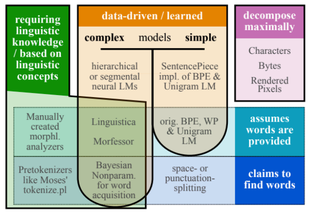


Figure 1: A taxonomy of segmentation and tokenization

Masader: Metadata Sourcing for Arabic Text and Speech Data Resources

Zaid Alyafeai[1], Maraim Masoud[2], Mustafa Ghaleb[1], and Maged S. Al-shaibani[1]

[1] King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia
[2] Independent Researcher

Abstract

The NLP pipeline has evolved dramatically in the last few years. The first step in the pipeline is to find suitable annotated datasets to evaluate the tasks we are trying to solve. Unfortunately, most of the published datasets lack metadata annotations that describe their attributes. Not to mention, the absence of a public catalogue that indexes all the publicly available datasets related to specific regions or languages. When we consider low-resource dialectical languages, for example, this issue becomes more prominent. In this paper we create *Masader*, the largest public catalogue for and so on. This study attempts to identify the publicly available Arabic NLP datasets and to provide a catalogue of Arabic datasets to researchers. The catalogue will increase the discoverability and provide some key metadata that will help researchers identify the most suitable dataset for their research questions.

We highlight our contributions as the following:

- We create the largest catalogue with 25 attributes for 200 Arabic NLP and speech datasets.
- We design a metadata schema for annotating

Published as a conference paper at ICLR 2022

MULTITASK PROMPTED TRAINING ENABLES ZERO-SHOT TASK GENERALIZATION

Victor Sanh* Hugging Face
Albert Webson* Brown University
Colin Raffel* Hugging Face
Stephen H. Bach* Brown & Snorkel AI

Lintang Sutawika BigScience
Zaid Alyafeai KFUPM
Antoine Chaffin IRISA & IMATAG
Arnaud Stiegler Hyperscience
Teven Le Scao Hugging Face
Arun Raja I²R, Singapore
Manan Dey SAP
M Saiful Bari NTU, Singapore
Canwen Xu UCSD & Hugging Face
Urmish Thakker SambaNova Systems
Shanya Sharma Walmart Labs
Eliza Szczechla BigScience
Taewoon Kim VU Amsterdam
Gunjan Chhablani BigScience
Nihal V. Nayak Brown University
Debajyoti Datta University of Virginia
Jonathan Chang ASUS
Mike Tian-Jian Jiang ZEALS, Japan
Han Wang NYU
Matteo Manica IBM Research
Sheng Shen UC Berkeley
Zheng-Xin Yong Brown University
Harshit Pandey BigScience
Michael McKenna Parity
Rachel Bawden Inria, France
Thomas Wang Inria, France
Trishala Neeraj BigScience
Jos Rozen Naver Labs Europe
Abheesht Sharma BITS Pilani, India
Andrea Santilli University of Rome
Thibault Fevry BigScience
Jason Alan Fries Stanford & Snorkel AI
Ryan Teehan Charles River Analytics
Tali Bers Brown University
Stella Biderman Booz Allen & EleutherAI
Leo Gao EleutherAI
Thomas Wolf Hugging Face
Alexander M. Rush Hugging Face

ABSTRACT

Large language models have recently been shown to attain reasonable zero-shot generalization on a diverse set of tasks (Brown et al., 2020). It has been hypothesized that this is a consequence of implicit multitask learning in language models' pretraining (Radford et al., 2019). Can zero-shot generalization instead be directly induced by *explicit* multitask learning? To test this question at scale, we develop a system for easily mapping any natural language tasks into a human-readable prompted form. We convert a large set of supervised datasets, each with multiple prompts with diverse wording. These prompted datasets allow for benchmarking

Already 13 papers published or submitted

PromptSource: An Integrated Development Environment and Repository for Natural Language Prompts

Stephen H. Bach[*1,2] **Victor Sanh**[*3] **Zheng-Xin Yong**[1] **Albert Webson**[1] **Colin Raffel**[3]
Nihal V. Nayak[1] **Abheesht Sharma**[4] **Taewoon Kim**[5] **M Saiful Bari**[6] **Thibault Fevry**[7]
Zaid Alyafeai[8] **Manan Dey**[9] **Andrea Santilli**[10] **Zhiqing Sun**[11] **Srulik Ben-David**[12]
Canwen Xu[13] **Gunjan Chhablani**[7] **Han Wang**[14] **Jason Alan Fries**[15,2]
Maged S. Al-shaibani[8] **Shanya Sharma**[16] **Urmish Thakker**[17] **Khalid Almubarak**[18]
Xiangru Tang[19] **Dragomir Radev**[19] **Mike Tian-Jian Jiang**[20] **Alexander M. Rush**[3]
[1] Brown University [2] Snorkel AI [3] Hugging Face [4] BITS Pilani [5] VU Amsterdam
[6] NTU [7] BigScience [8] KFUPM [9] SAP [10] University of Rome [11] CMU [12] Technion
[13] UCSD [14] NYU [15] Stanford University [16] Walmart Labs [17] SambaNova Systems
[18] PSAU [19] Yale University [20] ZEALS [*] Equal Contribution

Abstract

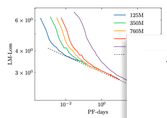
PromptSource is a system for creating, sharing, and using natural language prompts. Prompts are functions that map an example e.g. by mapping a response such as *"sports"* to a label class. In specific contexts, prompting has been shown to have advantages over traditional classification, for example facilitating adaptation

What Language Model to Train if You Have One Million GPU Hours?

Anonymous ACL submission

Abstract

The crystallization of modeling methods around the Transformer architecture has been a boon for practitioners. Simple, well-motivated architectural variations may transfer across tasks and scale, increasing the impact and leverage of modeling research. However, with the emergence of state-of-the-art 100B+ parameters models, large language models are increasingly expensive to accurately design and train. Notably, it can be difficult to evaluate how modeling decisions may impact emergent capabilities, given that these capabilities



Adapting BigScience Multilingual Model to Unseen Languages

Anonymous ACL submission

Abstract

We benchmark different strategies of adding new languages (German and Korean) into the BigScience's pretrained multilingual language model with 1.3 billion parameters that currently supports 13 languages. We investigate the factors that affect the language adaptability of the model and the trade-offs between computational costs and expected performance.

1 Introduction

is to benefit from knowledge transfer encoded in the pretrained LM for the new language processing at a small computational cost (compared to full model retraining) .

In this work, we aim at better understanding the trade-offs between the amount of compute and the final downstream task performance. Specifically, we study the impact of the following three factors: original pretraining steps, adaptation strategies, and adapters' capacity on the Natural Language Inference (NLI) task. As part of the initia-

And many other currently under blind submission

BigScience

The BigScience corpus

BigScience How did we create a 1,5 TB multilingual dataset?

INPUT FROM WGs

A/ Sourcing high-quality multilingual data

One of the training components is text extracted from the Catalog (output of the BigScience Data Sourcing Hackathon that resulted in 246 data resources).

B/ Identifying seeds from a web crawl

One of the training components is text extracted from a pseudo crawl. We initially identified seeds (605) from a web crawl to do so. We effectively retrieved text from 535 sources.

C/ Defining filters to apply to a web crawl

One of the components of the training set is text extracted from web crawl (OSCAR v2)

DATASET SPRINT FROM 01/28 TO 03/10

1 RETRIEVING

👉 We had a Hackathon on December 2021, during which contributors sourced the data from the Catalog and added it to the Hugging Face Hub.

👉 We extrated text via a loading script from the downloads and new datasets were loaded to the Hub. Many of the initial datasets were split into many datasets, since we created datasets per languages.

2 PREPROCESSING

👉 We examined individual sources to remove remaining pre-processing artifacts.

👉 We performed source-level line deduplication on selected datasets.

👉 We filtered items by length on higher-resourced languages.

3 LOADING DATA @HF

✅ Training data loadable as HuggingFace datasets.

1 RETRIEVING

👉 We created an index using the identified seeds in Common Crawl.

👉 We queried the index to retrieve WARC files and extracted the web pages (HTML format) from the WARC files.

👉 We extracted the text content from HTML web pages.

2 PREPROCESSING

👉 We performed URL-based deduplication.

👉 We performed a seed level line deduplication.

👉 We selected high priority filters (cf step C ▶️) to remove some pages :
- Length,
- Character repetition,
- Language ID confidence,
- Common token ratios.

3 LOADING DATA @HF

✅ Training data loadable as HuggingFace datasets.

1 RETRIEVING

👉 We downloaded OSCAR v2.

👉 To ensure that humans wrote the retrieved text for humans, it required the creation of filters to exclude the "spam" pages from OSCAR v2. We collected inputs from native speakers for non-language agnostics filters (flagged word ratio, closed class words ratio). We created a tool to manage filtering thresholds for all the other filters.

👉 We were able to retrieve text from 13 languages.

2 PREPROCESSING

👉 We performed deduplication.

👉 We used 13 filters to remove "spam" pages: Length, character repetition, Language ID, Stopwords, flagged word ratio...

👉 We removed several categories of PII (Personally identifiable information): email, username, IP address...

3 LOADING DATA @HF

✅ Training data loadable as HuggingFace datasets.

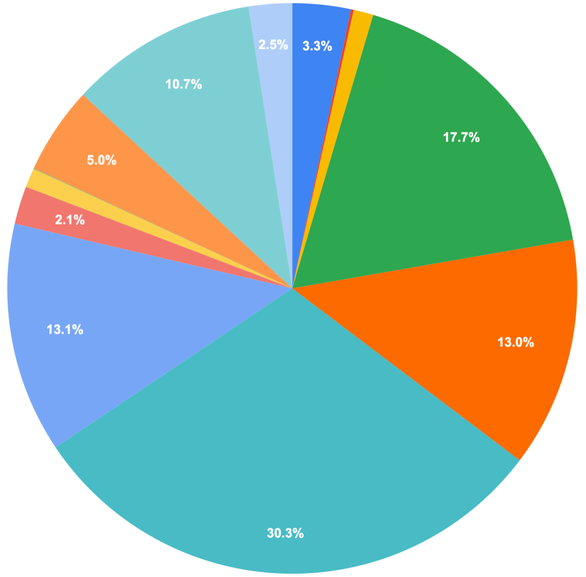
We collected text data from a human-curated catalog of data sources in all BigScience languages to best leverage our language expertise for dataset quality. This step still needed to be complemented with other approaches to meet our scale and diversity requirement, so we also used pseudo-crawled (column B) and crawled (column C) data.

Our pseudo-crawl data was made up of specific websites selected by participants to maximize geographical diversity, especially for English and Spanish-language data. This data required additional filters to handle the noise and artifacts of web content.

Catalog (column A) and Pseudo-Crawl (column B) data together accounted for 65% of our target corpus size while still over-representing English. We complemented it with data obtained from a pre-existing web crawl (OSCAR v2) to improve the diversity and balance of the final dataset.

DATASET

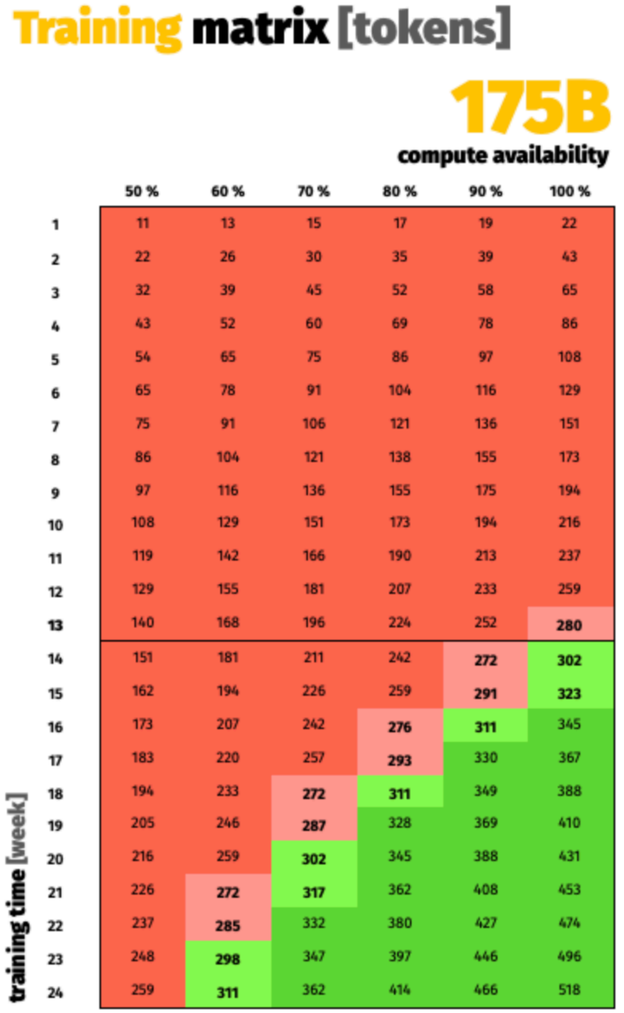
350B tokens (1.5 TB) multilingual dataset



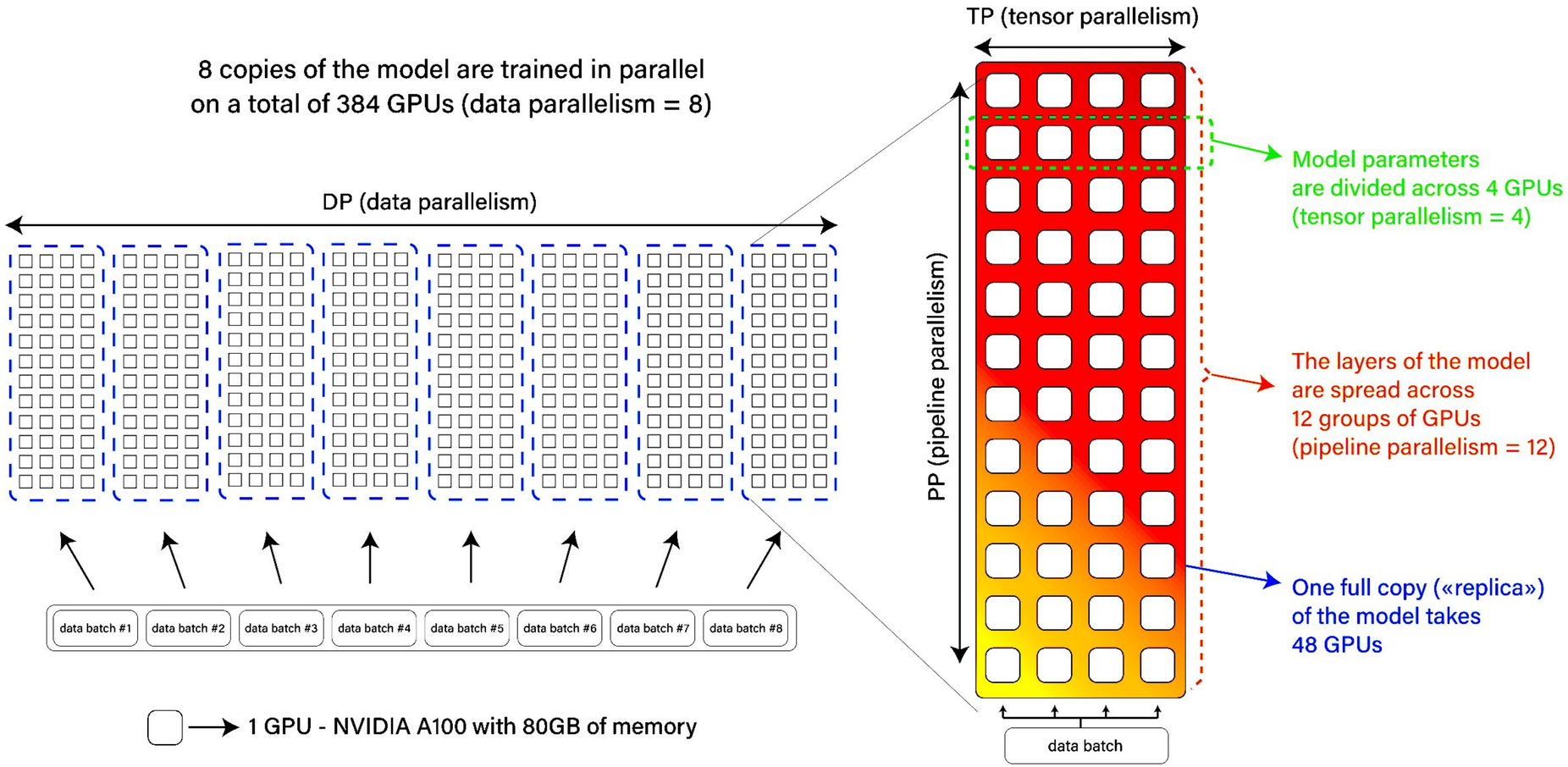
- Arabic (3,3%)
- Basque (0,2%)
- Catalan (1,1%)
- Chinese (17,7%)
- Code (13%)
- English (30,3%)
- French (13,1%)
- Indic (2,1%)
- Indonesian (1,1%)
- Niger Congo (0,03%)
- Portuguese (5%)
- Spanish (10,7%)
- Vietnamese (2,5%)

BigScience

- **The multilingual 176B parameters Language Model**
  - **176B parameters** decoder transformers
  - 70 layers - 112 attention heads - 14336 hidden dimensionality- 2048 tokens per sequence
  - ALiBi positional embeddings - GeLU activation
- **Hardware**:
  - **384 GPUs NVIDIA A100 80GB**
  - Weights size: 329GB
  - Throughput: 150 TFLOPs
- **Environment**:
  - Cluster mostly powered by nuclear energy
  - Heat generated used for heating campus
- **All information open** on the BigScience website



BigScience



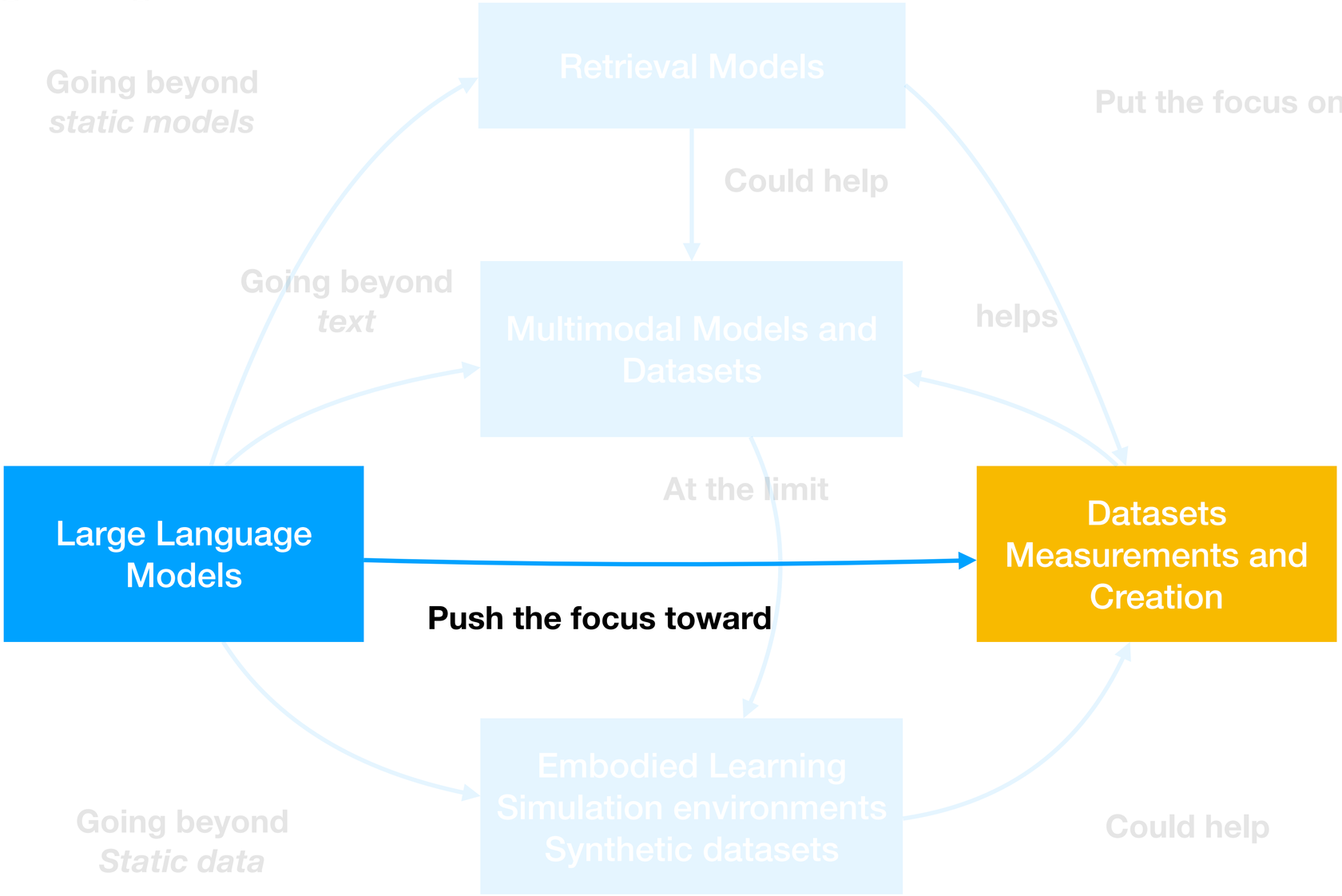
Current status: **13% of the training**

Expected end of training date: End of June

BigScience

- **To follow or join**
  - Website: https://bigscience.huggingface.co
  - Twitter: @BigScienceW (general workshop)
    @BigScienceLLM (176B parameters model training)
- **What's next / meet the people**
  - May 27th – ACL 2022 Workshop – BigScience Episode #5
    "Challenges & Perspectives in Creating Large Language Models" *(21 submissions)*

Hugging Face current research directions



Datasets

- Focus on the models but
  - Today's architectures are **very simple**
    (transformers very close to the original model)
  - Models are only a **reflection** of their training data
  - The real source of most recent progress: increase in **data size and diversity**
- What do we know about our data today?

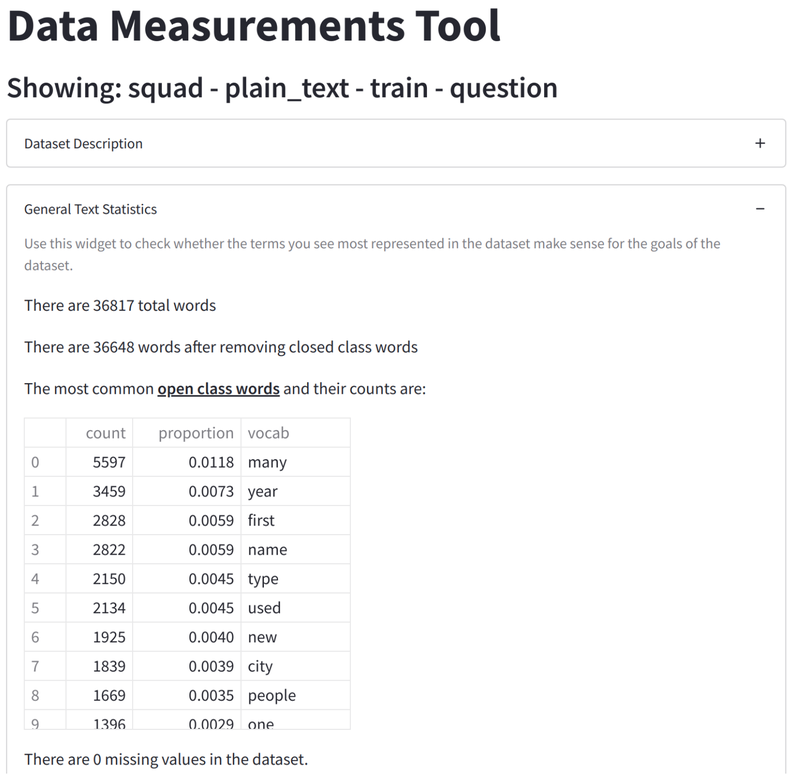
Datasets

- We need to
  - **Understand/measure** datasets better – bias, diversity, quality, representativeness, etc
  - **Get better at building** datasets from ethical/responsible/legal/ML point of view

- A first step:
  the **Data Measurements Tool**

  A tool to explore datasets with aggregated and fine measurements

  - perplexity, bias, duplicates, stats, zip law, many data metrics



Data Measurements Tool

Showing: squad - plain_text - train - question

Dataset Description +

General Text Statistics −

Use this widget to check whether the terms you see most represented in the dataset make sense for the goals of the dataset.

There are 36817 total words

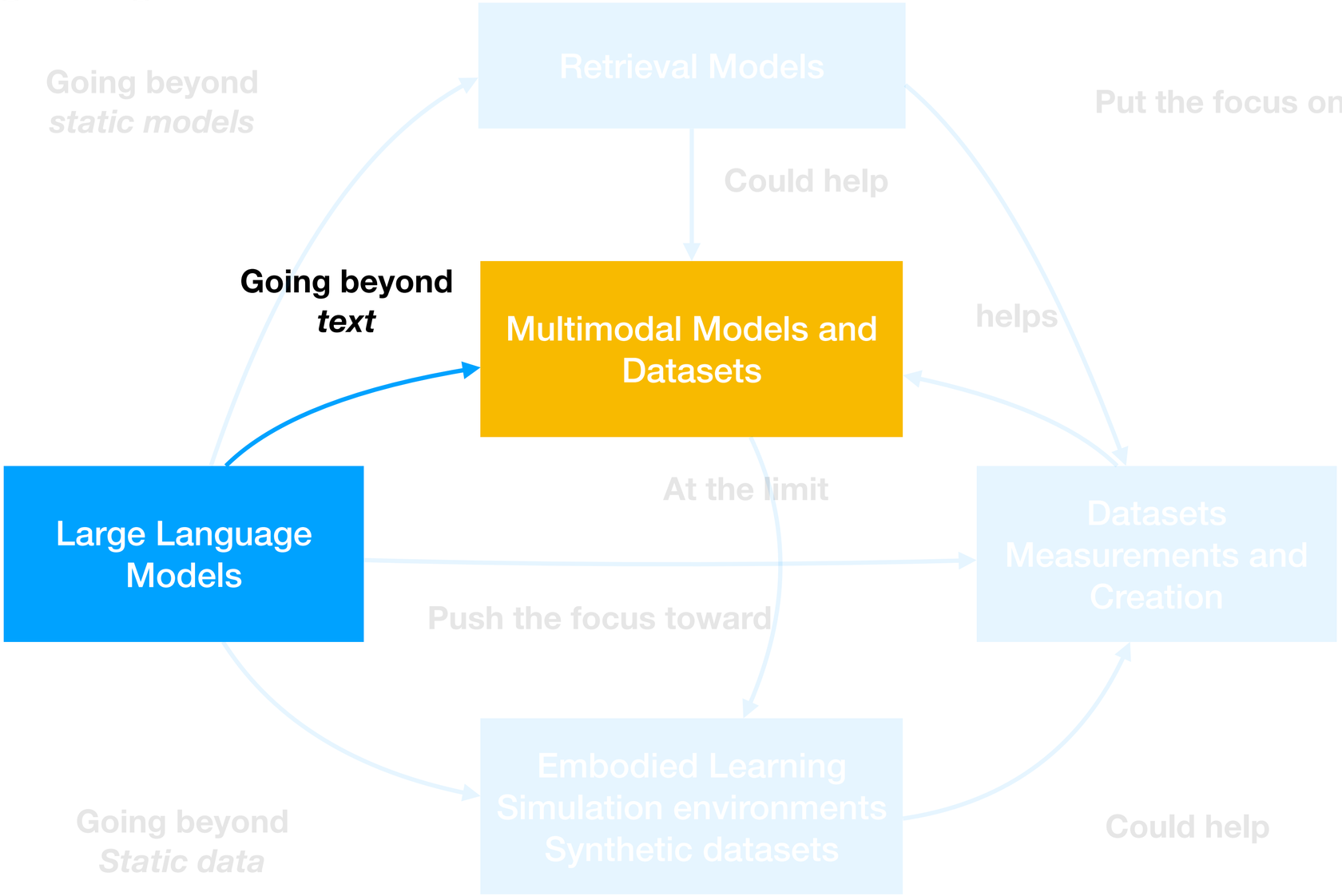
There are 36648 words after removing closed class words

The most common **open class words** and their counts are:

	count	proportion	vocab
0	5597	0.0118	many
1	3459	0.0073	year
2	2828	0.0059	first
3	2822	0.0059	name
4	2150	0.0045	type
5	2134	0.0045	used
6	1925	0.0040	new
7	1839	0.0039	city
8	1669	0.0035	people
9	1396	0.0029	one

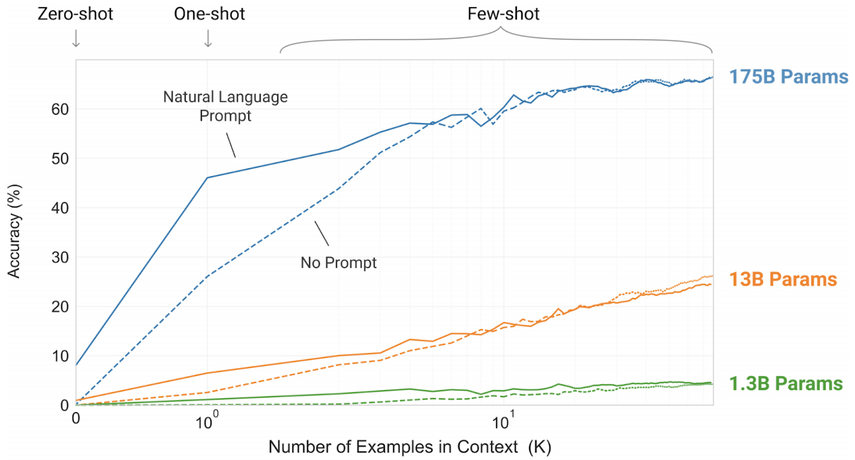
There are 0 missing values in the dataset.

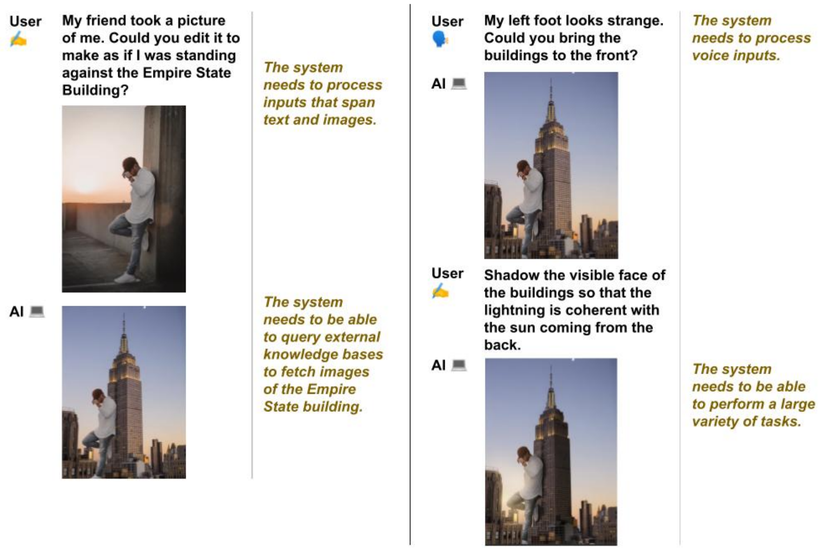
Hugging Face current research directions



Multi-modality

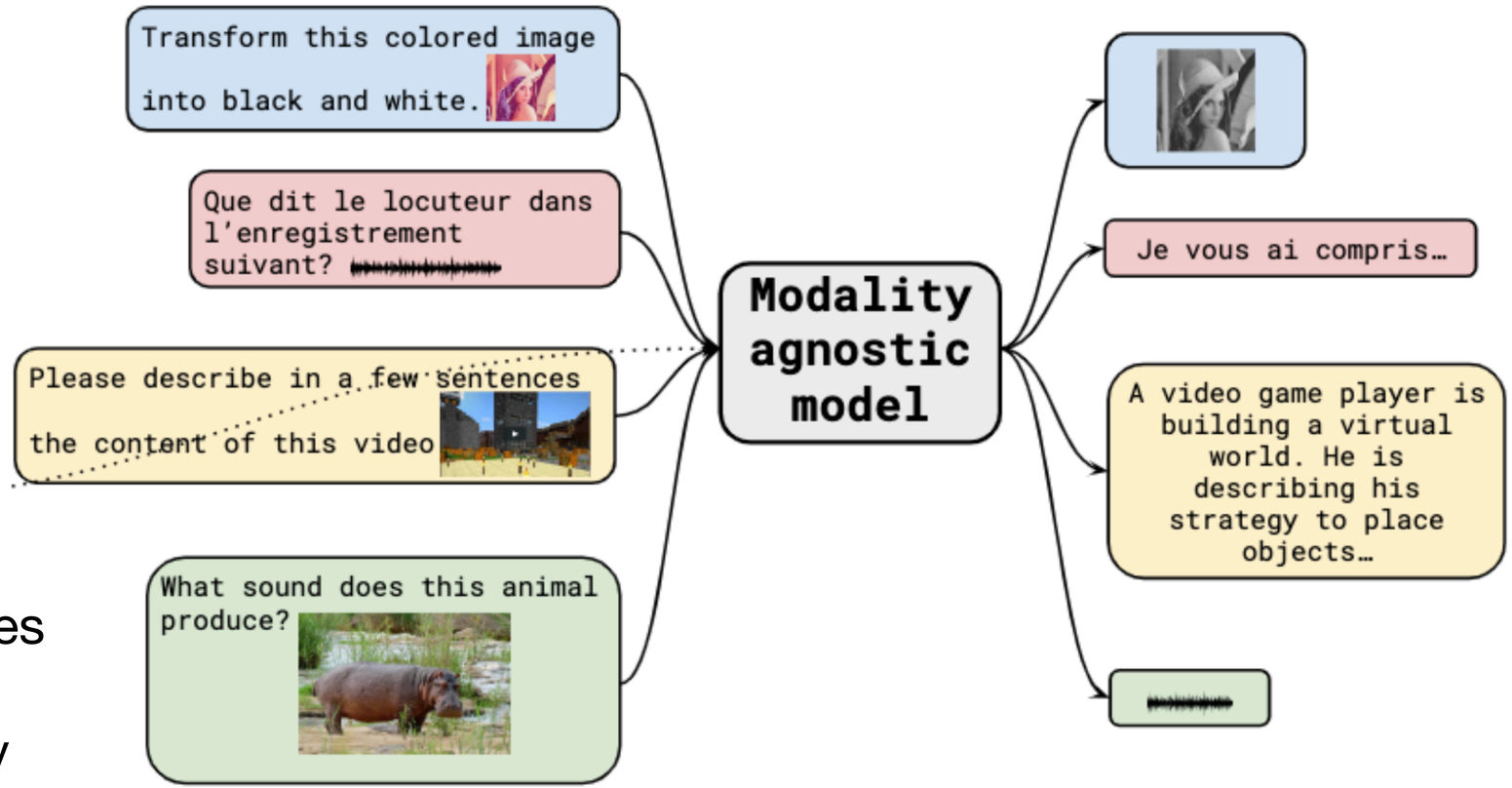
- Why going multi-modal?
  - Scaling laws means decreasing returns
  - Hard/not-data-efficient to learn some aspects of world knowledge from text only (visual concepts / reporting bias)
  - Unlocking new applications and new model behaviors
  - Toward embodied AI



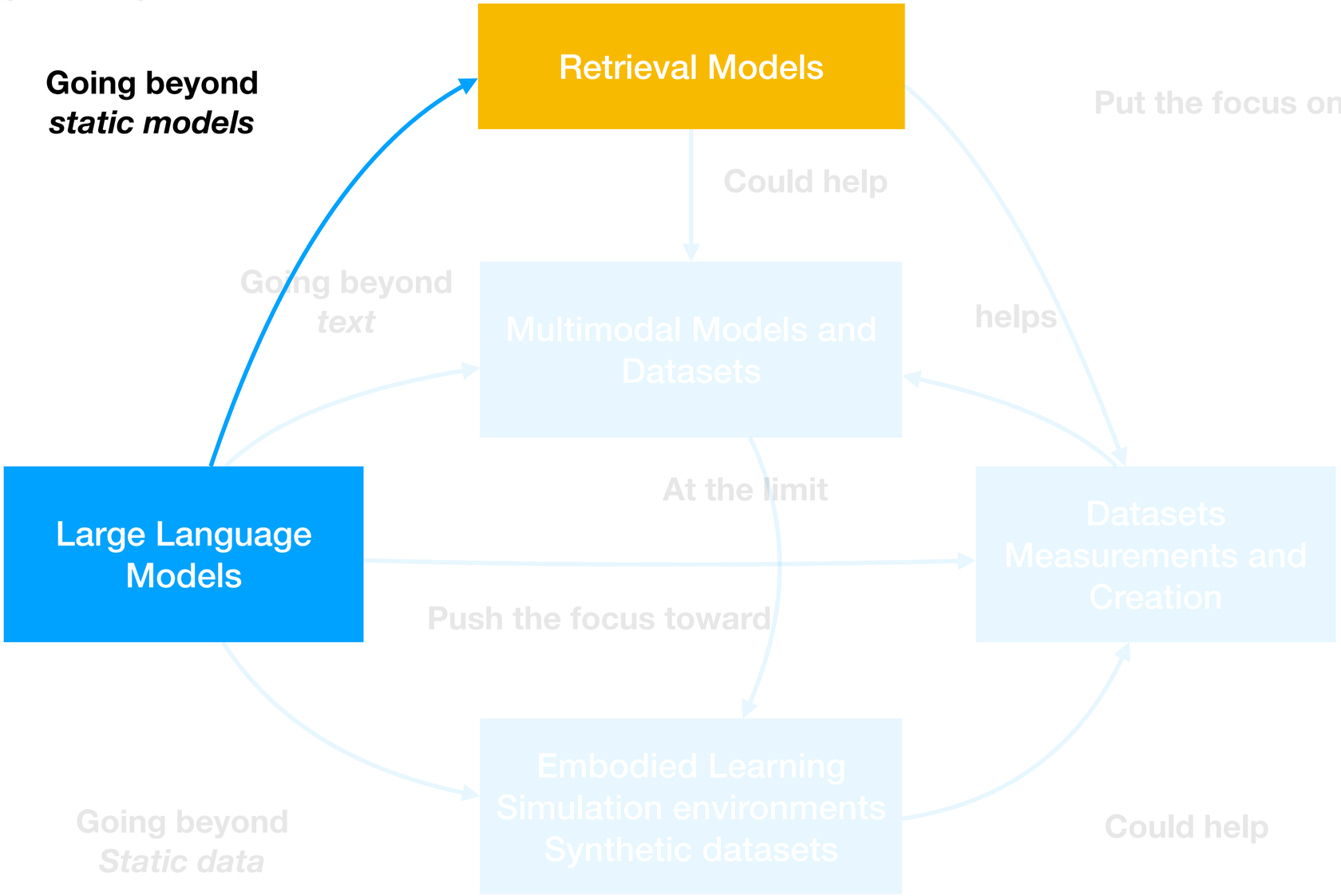


Multi-modality

- Challenges and opportunities
  - Data
    - Which modalities?
    - Data availability?
  - Models
    - Universal compute engines
    - The question of efficiency
    - Futuristic architectures models

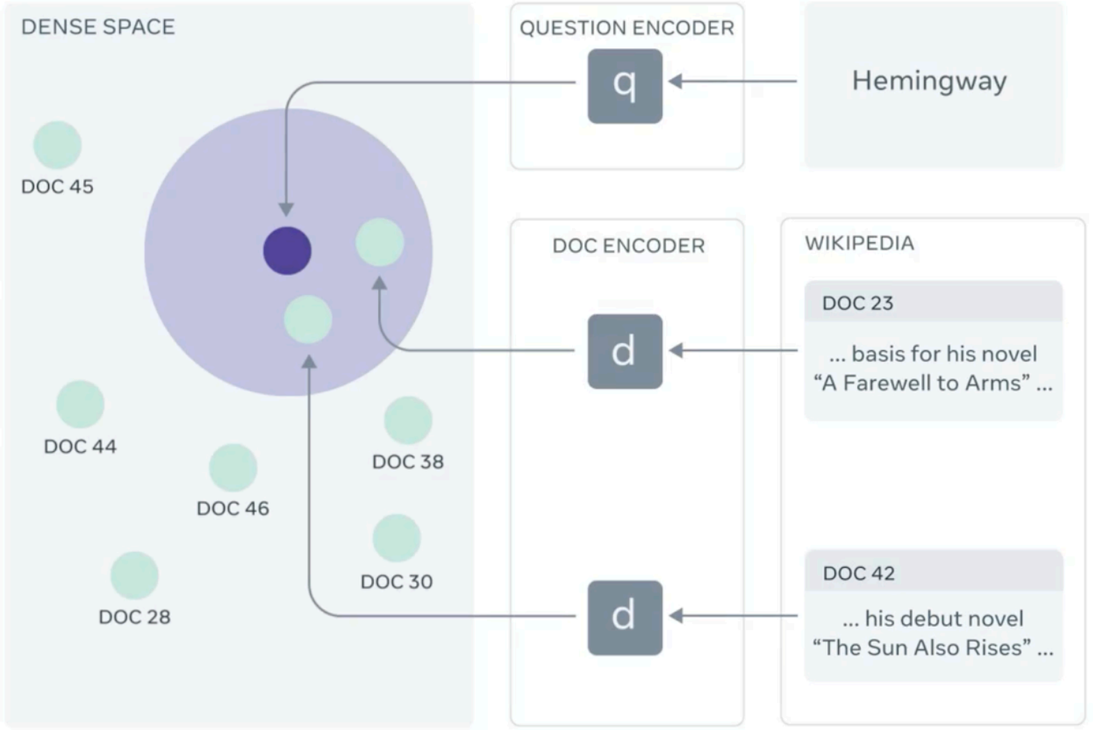


Hugging Face current research directions



Retrieval models

- Overcoming the **limits of static models**
  - BERT and the president of the USA
  - GPT3 and COVID-19
- Retrieval as a way to **decouple knowledge** from **computation**



Retrieval Augmented Generation – Lewis et al

Retrieval models

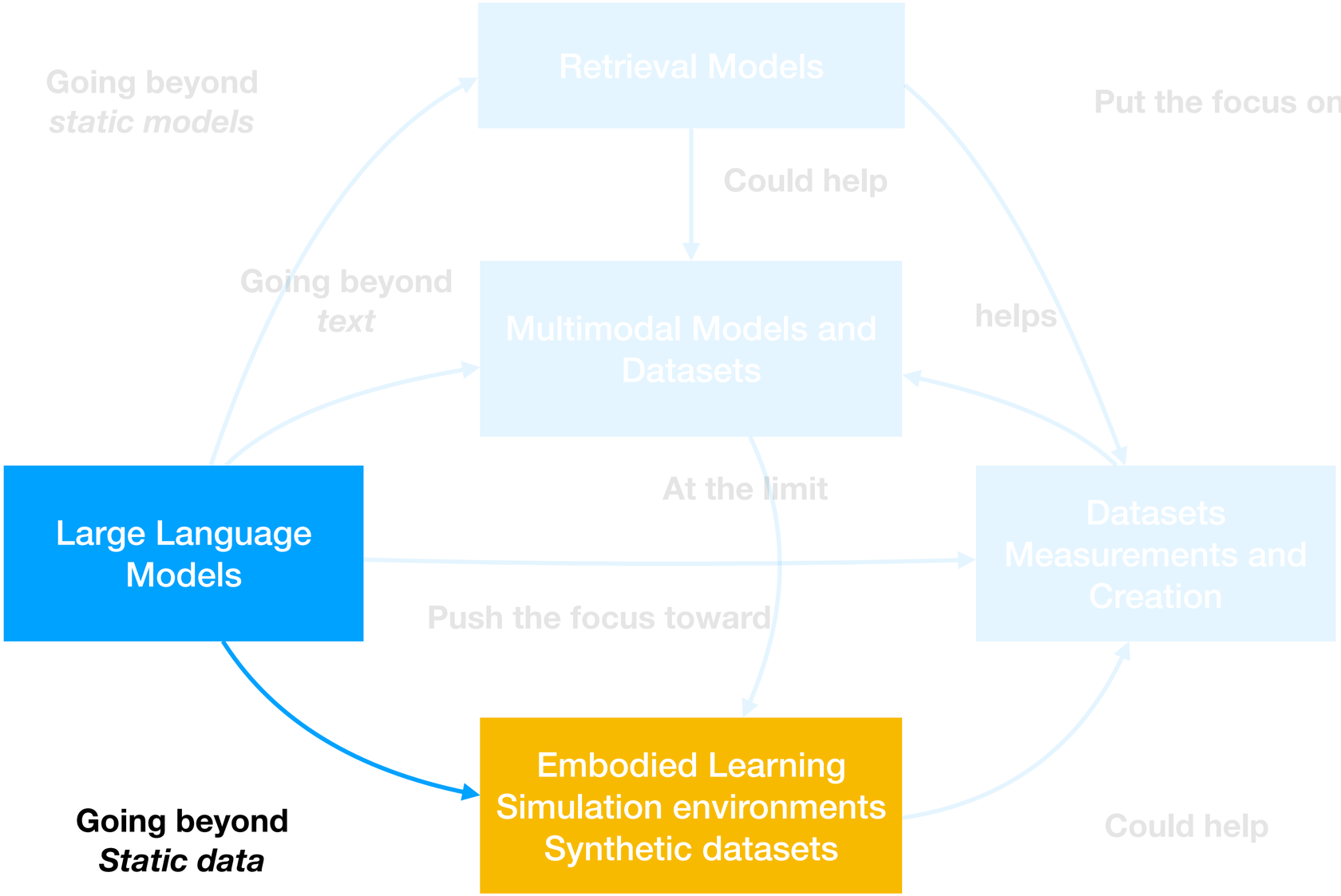
➡ **Multilingual Dense Retrieval Model**

- **Non-English** languages: *performance a lot weaker than English for semantic search*
- **Lack of training** data => Create large scale (2B+) dataset

➡ **Few-Shot Text Classification** based on **vector spaces**

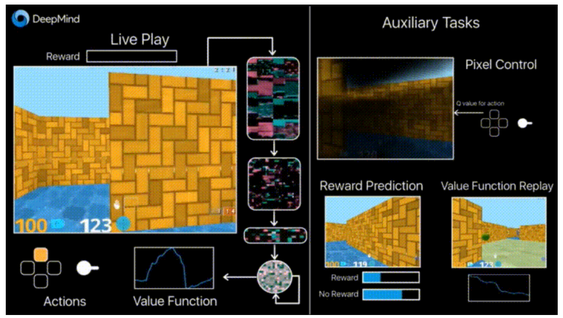
➡ **Inject new knowledge** into existing models

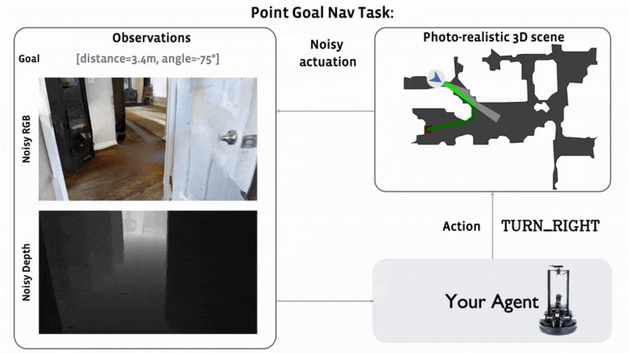
Hugging Face current research directions

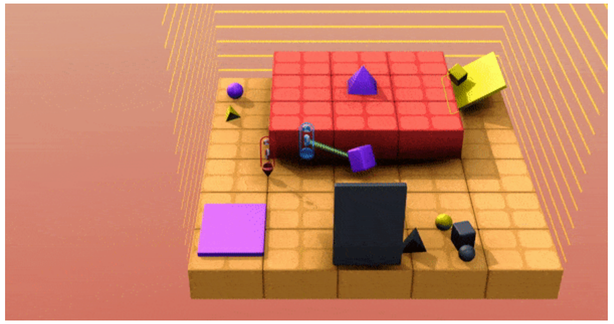


Embodied learning

- Going toward models that **learn from interactions** & diverse inputs
- Simulation environnements are starting see **scaling laws** as well
- But difficult they are to use, to share, to investigate, to reproduce
- Maybe Hugging Face can help :-)

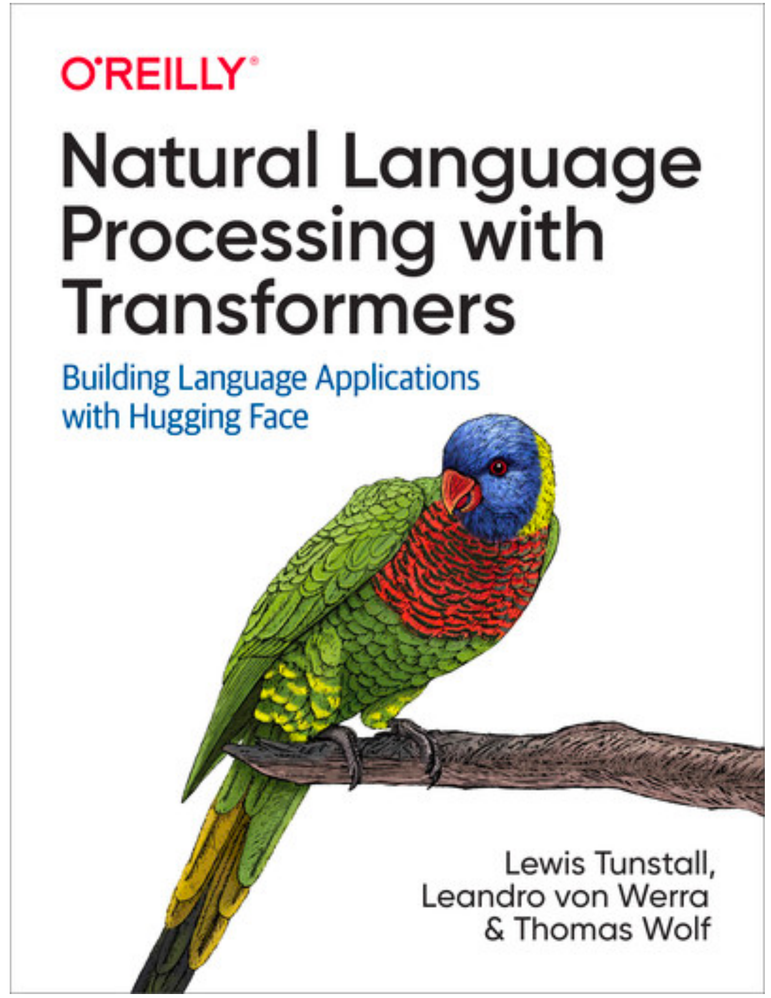






One last thing

- We have a book just published in 2022 at O'Reilly!
- Covers all about using transformers in practical NLP applications (classification, summarization, use in production…)
- I brought **20 free copies** that I'll drop on the **badge table** later today
- Grab them!



The End – Thanks a lot for your attention!