TReNDS
Translational Research in
Neuroimaging & Data Science

Georgia Tech
College of
Engineering

Galactica: A large language model for science

By Taylor et al. (Meta AI), 2022

Spring 2023 **02/17/2023**

Summary

- Introduction
- Dataset & Methods
- Results
- Further analysis

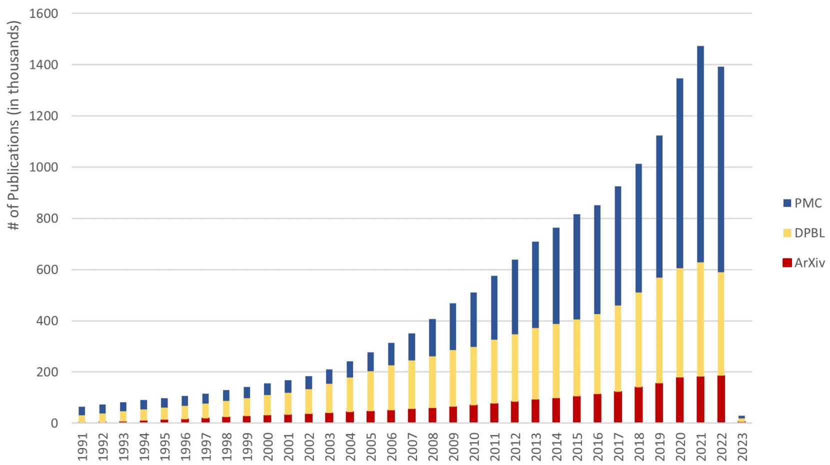
Summary

Information overload: a long time predicted burden

A problem already known decades ago…

« Publication has been extended far beyond our present ability to make real use of the record »

Vannevar Bush, *As We May Think*, 1945



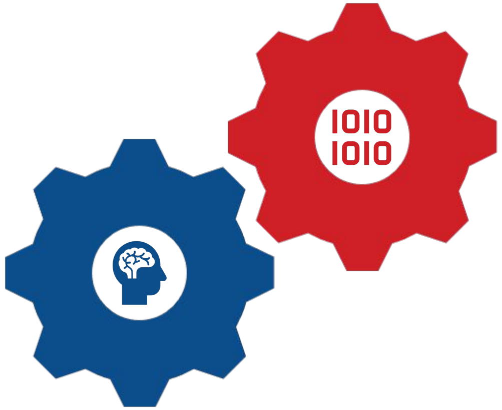
… reaching a point of no return

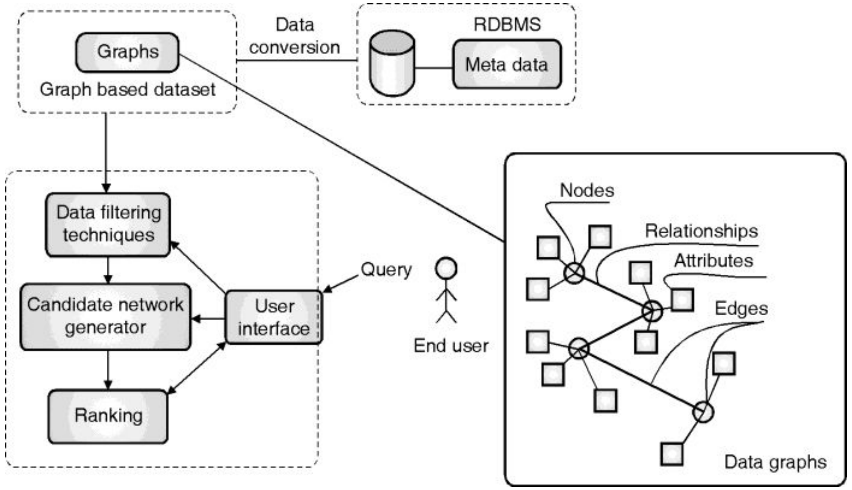
- A publication rate way above the capabilities of scientists to read them: an average of 516 publications submitted per day on ArXiv (May 2022)
- An overload not only limited to publication: for instance, NCBI GenBank contained almost 1.5×10^{12} nucleotide bases in August 2022

Technology: a potential solution

Computers, instrument to reach Licklider's paradigm

- The rise of information technology and computers : invention of the transistor in 1947 (by Bardeen, Shockley and Brattain), of microprogramming in 1955 (by Maurice Wilkes)…
- A source of hope to tackle the issue: in Licklider's paradigm, computer would *"prepare the way for insights and decisions in scientific thinking"* (Licklider, *Man-Computer Symbiosis*, 1960)





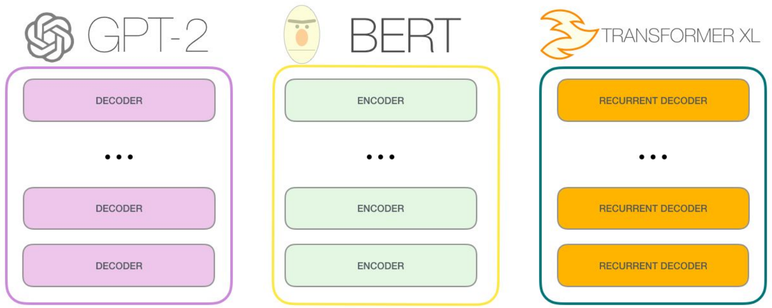
… but still needing too much human contributions

- The current "symbiotic" relationship between human and computer still need a lot of human contribution when information need to be found (search engines)
- A task, even with the use of computers, that is still time-consuming

Large Language Models: a breakthrough in NLP

Large Language Models (LLM)

- LLMs have achieved breakthrough performance on NLP tasks in last year.
- Some argue that Language Models can be considered as a convenient implicit knowledge bases



Galactica: a new LLM for organizing science

- A dataset of more than 48 millions papers, textbooks… but also proteins, DNA sequences…
- A particular focus on the dataset, « high-quality and highly curated »
- A model that beat previous LM on several benchmarks (MMLU, MATH…)

Summary

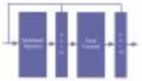
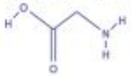
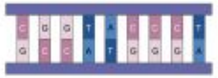
- Introduction
- **Dataset & Methods**
- Results
- Further analysis

Galactica’s dataset: heart of the model (1/2)

A large scientific corpus

- More than 60 million documents coming from 6 main data source used to train Galactica
- All document converted in Markdown to unify knowledge coming from all kind of documents
- Text sequence only, but many scientific phenomena described

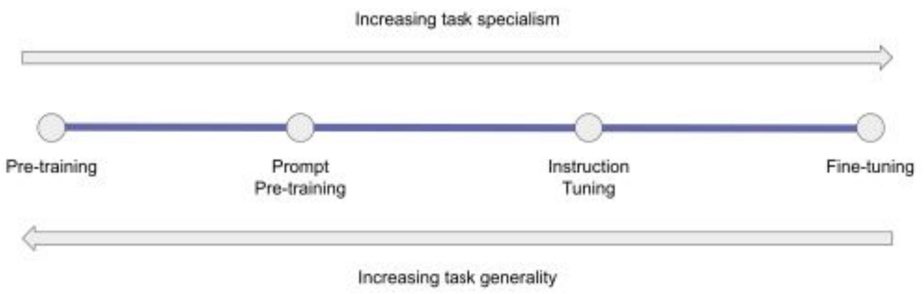
Total dataset size = 106 billion tokens			
Data source	**Documents**	**Tokens**	**Token %**
Papers	48 million	88 billion	83.0%
Code	2 million	7 billion	6.9%
Reference Material	8 million	7 billion	6.5%
Knowledge Bases	2 million	2 billion	2.0%
Filtered CommonCrawl	0.9 million	1 billion	1.0%
Prompts	1.3 million	0.4 billion	0.3%
Other	0.02 million	0.2 billion	0.2%

Modality	Entity	Sequence	
Text	Abell 370	`Abell 370 is a cluster...`	
LaTeX	Schwarzschild radius	`r_{s} = \frac{2GM}{c^2}`	$r_s = \frac{2GM}{c^2}$
Code	Transformer	`class Transformer(nn.Module)`	
SMILES	Glycine	`C(C(=O)O)N`	
AA Sequence	Collagen α-1(II) chain	`MIRLGAPQTL..`	
DNA Sequence	Human genome	`CGGTACCCTC..`	

Galactica's dataset: heart of the model (2/2)

Prompt Pre-Training

- PPT can boost performance (lower models beating larger ones on specifics tasks)
- Be able to gives correct performances even for the smallest version of the model
- Almost 800k prompts given on different tasks (summarization, entity extraction, binary QA…)
- PPT create a distinction between in-domain knowledge and out-domain knowledge



Tokenization: break data into understandale items

Specialized tokenization: a choice for the dataset design

Special type of data	Choice of tokenization
Step-by-step reasoning	Wrapping with <work>
Citations	Wrapping with [START_REF] / [END_REF]
SMILES formula, DNA sequences and Amino acid sequences	Wrapping with [START_SMILES] / [END_SMILES] ([START_DNA] / [END_DNA] or [START_AMINO] / [END_AMINO]) and character-based tokenization
Mathematics and numbers	Splitting digits and operations into individual characters

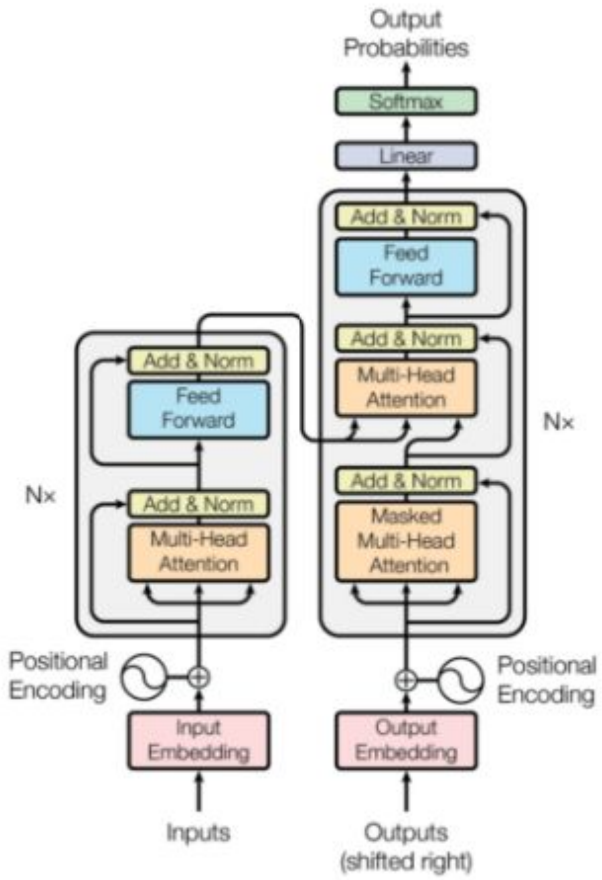
Recurrent neural networks, long short-term memory [START_REF] Long Short-Term Memory, Hochreiter [END_REF] and gated recurrent [START_REF] Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling, Chung [END_REF] neural networks in particular, have been firmly established as state of the art approaches in sequence modeling and transduction problems such as language modeling and machine translation [START_REF] Sequence to Sequence Learning with Neural Networks, Sutskever [END_REF] [START_REF] Neural Machine Translation by Jointly Learning to Align and Translate, Bahdanau [END_REF] [START_REF] Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation, Cho [END_REF].

Transformers: Galactica’s architecture (1/2)

Transformers : current state-of-the-art models

- Transformer architecture was introduced in June 2017, mainly to work on translation task
- Two main blocks: an encoder to receive inputs and build a representation of them, and a decoder using encoder representation and others inputs to generate a target sequence
- Each block can be used without the other, hence three main types of models.

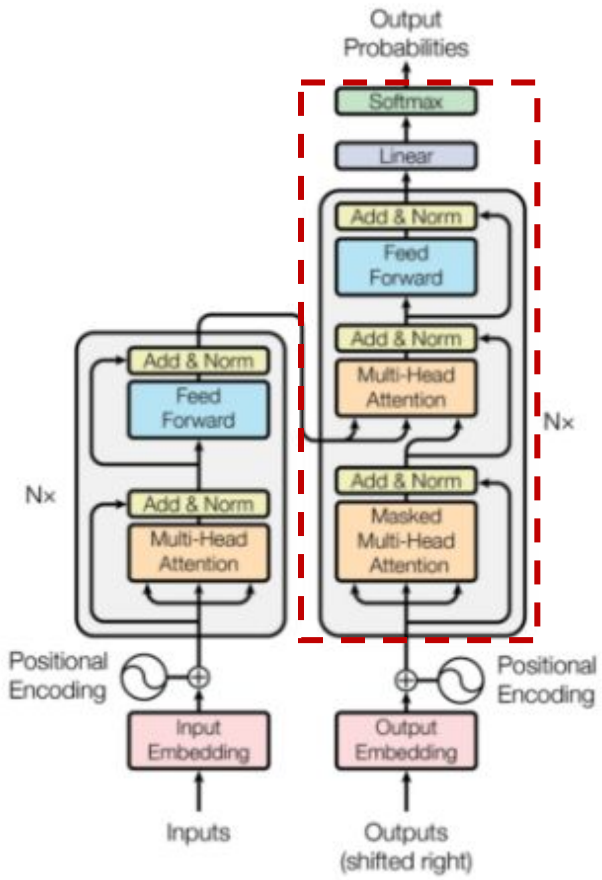
Models	Tasks	Exemple of models
Auto-Encoding	Sentence classification, NER	BERT
Auto-Regressive	Text Géneration	GPT
Sequence-to-Sequence	Translation, summarization	BART / T5



Transformers: Galactica’s architecture (2/2)

Galactica’s architecture : a modified version of the original architecture

- Only a decoder part (like GPT)
- Use of GELU activation function for all model in last feed forward layer
- No biases
- Use of Learned Position Embedding
- Creation of a 50k token vocabulary using BPE



<work> : a working memory token

A simple observation leading to this token

	Transformers		Classic computers
+	› Understanding of natural language › Chain-of-thought	+	› Arithmetic tasks
−	› Accuracy on task like multiplication	−	› Chain-of-thought

Question: A needle 35 mm long rests on a water surface at 20°C. What force over and above the needle's weight is required to lift the needle from contact with the water surface? $\sigma = 0.0728\text{m}$.

<work>

$$\begin{aligned} \sigma &= 0.0728\ \text{N/m} \\ \sigma &= F/L \\ 0.0728 &= F/(2 \times 0.035) \\ F &= 0.0728(2 \times 0.035) \end{aligned}$$

calculate.py

```
f = 0.0728*(2*0.035)

with open("output.txt", "w") as file:
    file.write(str(round(f, 5)))
```

«run: "calculate.py">

«read: "output.txt"»

0.0051

</work>

Answer: $F = 0.0051$ N

Process behind the creation of the token

Instruction → Single-forward pass → Detection of model limitations → Offloading

Summary

› Introduction

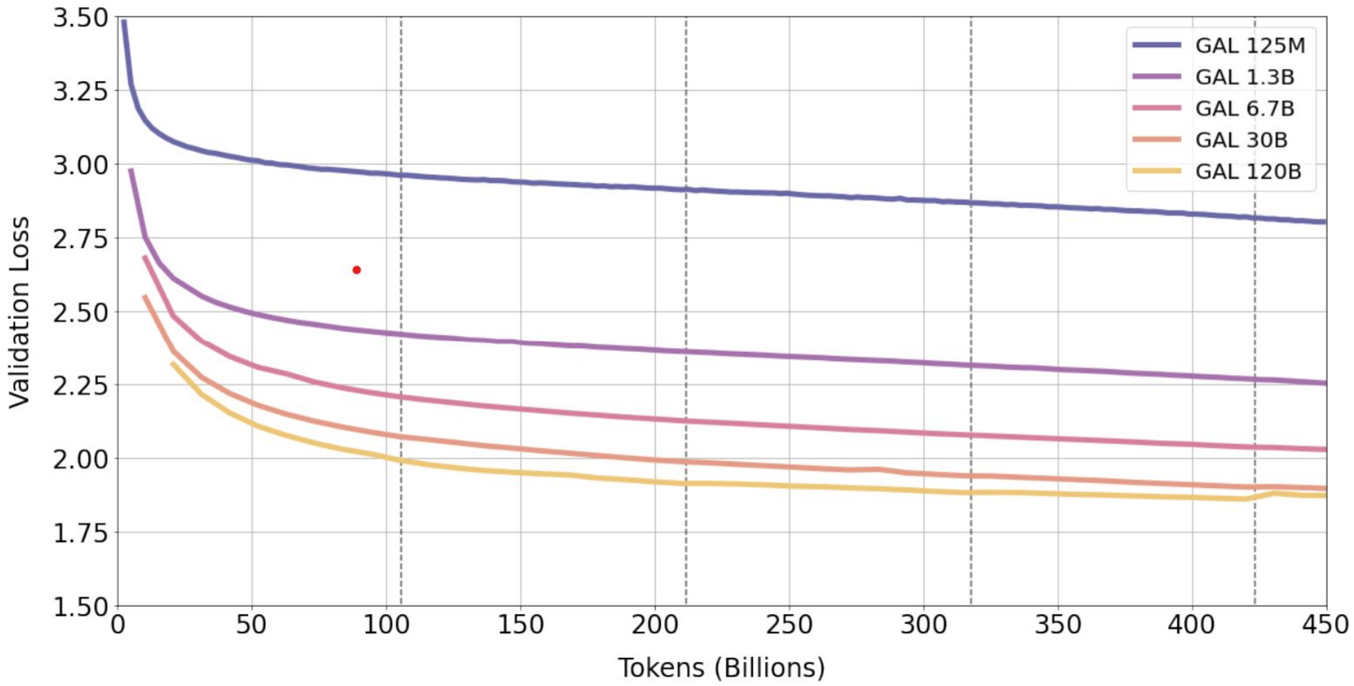
› Dataset & Methods

› **Results**

› Further analysis

Galactica models

Model	n_{params}	n_{layers}	d_{model}	n_{heads}	d_{heads}	Batch Size	Max LR	Warmup
GAL 125M	125M	12	768	12	64	0.5M	6×10^{-4}	375M
GAL 1.3B	1.3B	24	2,048	32	64	1.0M	2×10^{-4}	375M
GAL 6.7B	6.7B	32	4,096	32	128	2.0M	1.2×10^{-4}	375M
GAL 30B	30.0B	48	7,168	56	128	2.0M	1×10^{-4}	375M
GAL 120B	120.0B	96	10,240	80	128	2.0M	0.7×10^{-5}	1.125B



Results

Knowledge probes

Tasks	Galactica	Others models
LaTeX equations probes	68.2%	49% (GPT-3)
Domain probes	8 – 43.1%	9.7 – 35.1%
Reasoning	41.3%	35.7% (Chinchilla)

General capabilities

Model	Params (bn)	Accuracy *weighted*	Accuracy *unweighted*
OPT 30B	30	39.6%	38.0%
BLOOM 176B	176	42.6%	42.2%
OPT 175B	175	43.4%	42.6%
GAL 30B	30	46.6%	42.7%
GAL 120B	120	**48.7%**	**45.3%**

BIG-bench 57 task results

Downstream scientific NLP

	Galactica	Others models
In-domain	5	0
Out-domain	6	14

Numbers of dataset where models has best performance

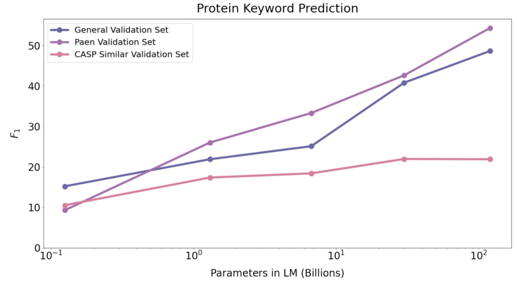
Chemical understanding

- **IUPAC Name Prediction**: accuracy of 39.2%
- **MoleculeNet**: Uni-Mol performs better

Citation prediction

Tasks	Galactica	Others models
PWC Citations	51.9%	30.9%
Extended Citations	69.1%	17.3%
Contextual Citations	36.6%	8.2%

Biological understanding



Summary

- Introduction
- Dataset & Methods
- Results
- **Further analysis**

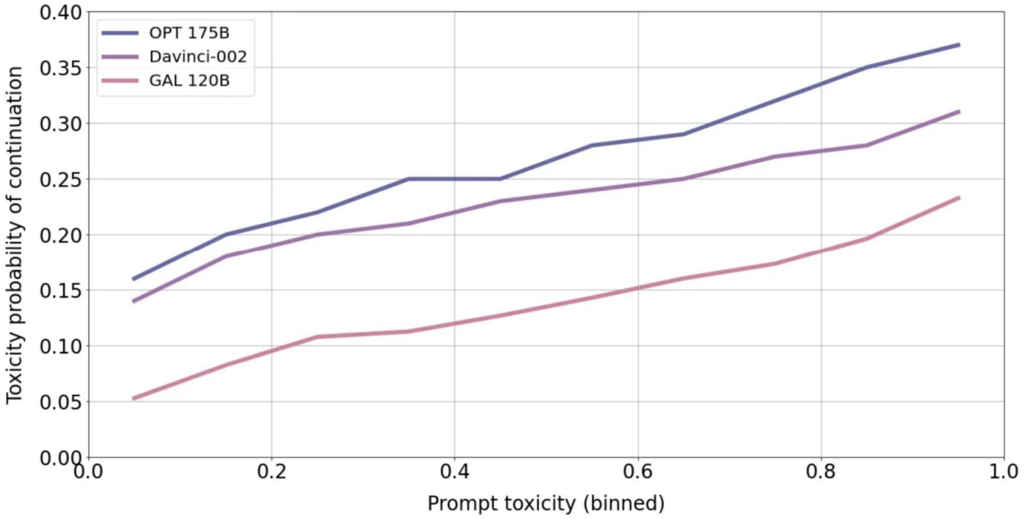
Toxicity and Bias

Generating content: a potential open door for toxicity

- Meta AI aware of the potential toxicity coming from LLM
- Use of benchmarks on toxicity and stereotypes to ensure Galatica's ability to detect stereotypes
- Galactica demo was shut down few days after its launch due many users retrieving biased, offensive or false answers to their questions

StereoSet				
Category		text-davinci-002	OPT 175B	Galactica 120B
Prof.	LMS (↑)	78.4	74.1	75.2
	SS (↓)	63.4	62.6	57.2
	ICAT (↑)	57.5	55.4	**64.3**
Gend.	LMS (↑)	75.6	74.0	74.6
	SS (↓)	66.5	63.6	59.1
	ICAT (↑)	50.6	53.8	**61.0**
Reli.	LMS (↑)	80.8	84.0	81.4
	SS (↓)	59.0	59.0	55.1
	ICAT (↑)	66.3	68.9	**73.1**
Race	LMS (↑)	77.0	74.9	74.5
	SS (↓)	57.4	56.8	54.8
	ICAT (↑)	65.7	64.8	**67.3**
Overall	LMS (↑)	77.6	74.8	75.0
	SS (↓)	60.8	59.9	56.2
	ICAT (↑)	60.8	60.0	**65.6**

StereoSet Results



Toxicity rate on RealToxicityPrompts

Limitations and Potential work

Limitations highlighted

- Limitations coming from corpus
- Distinguishbility of corpus effects and prompt effects
- Bias for highly-cited papers
- Text as only modality
- …

Several ideas mentionned

- Use of larger context window
- Extending to images
- Create more examples for the working memory token
- Enforce a verification layer
- Develop a continual learning
- …

TReNDS
Translational Research in
Neuroimaging & Data Science