

Possibilities and Limitations of Large Language Models: PAGnol, VLM-4 and Muse



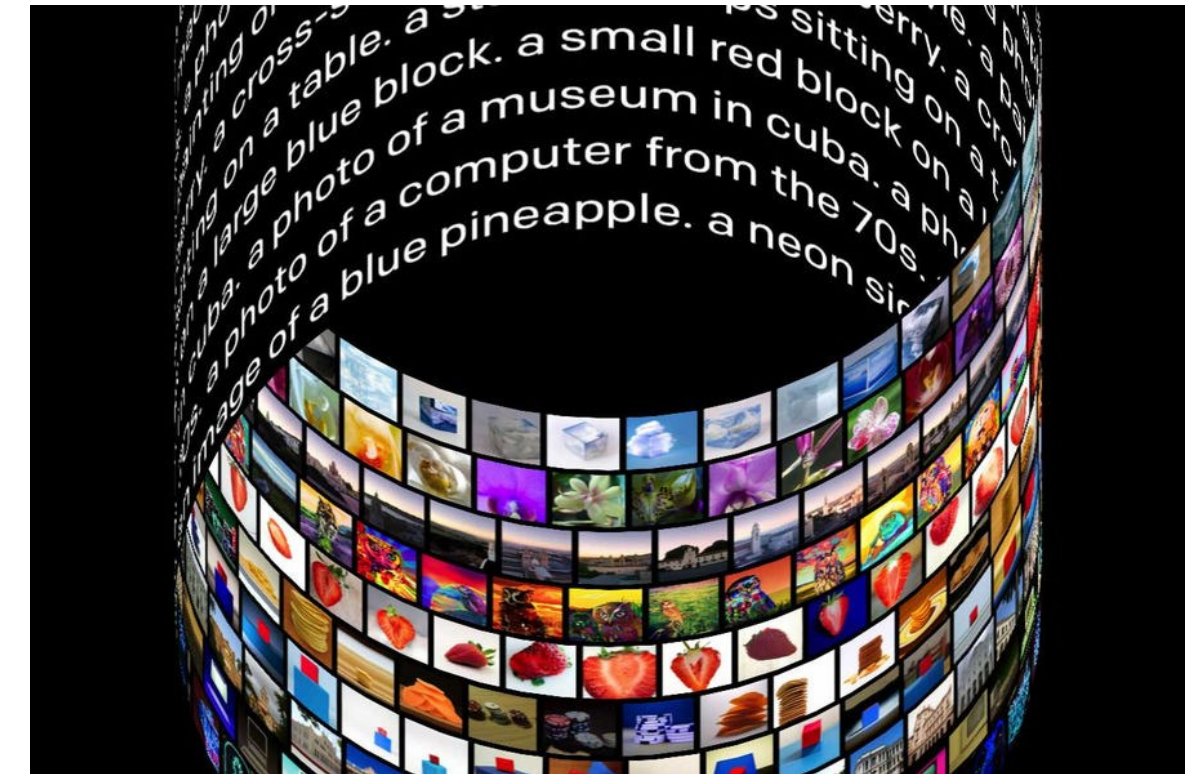
Igor Carron, PhD
LightOn, CEO
contact@lighton.ai



A **new generation of AI** is disrupting

Office work through text-based AI
Code programming
Drug discovery

GPT-3 (2020)
Codex (07/2021), AlphaCode (02/2022)
AlphaFold (08/2021)



With a **strong effect on technology and business** in record time



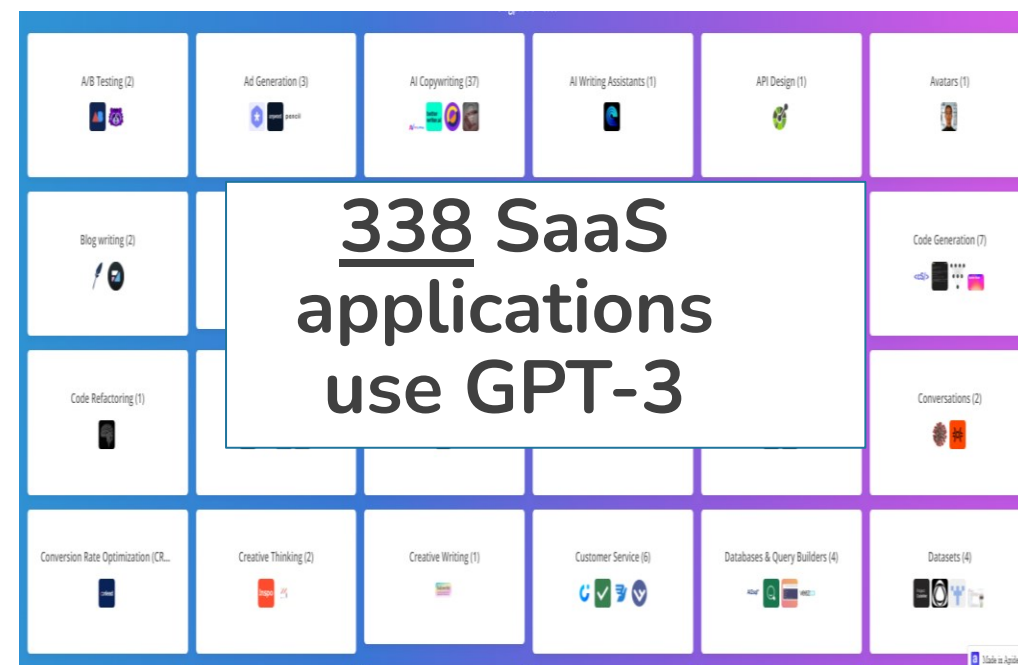
DeepMind's AlphaCode AI writes code at a competitive level

Devin Coldewey @t... 2, 2022

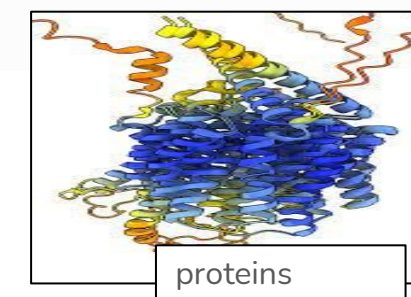


computer code

Comment



Forbes
AlphaFold Is The Most Important Achievement In AI—Ever

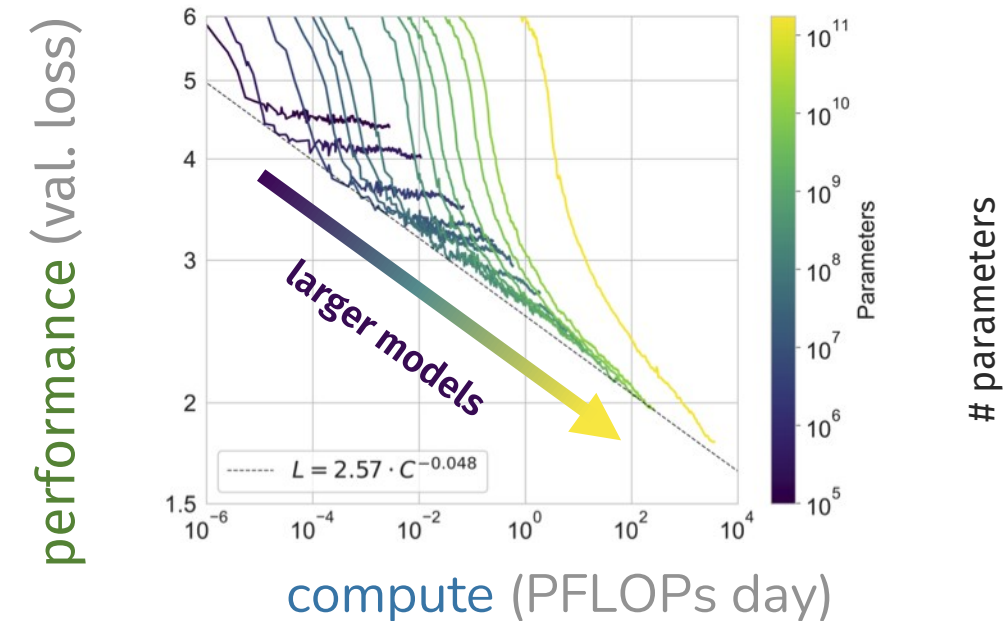
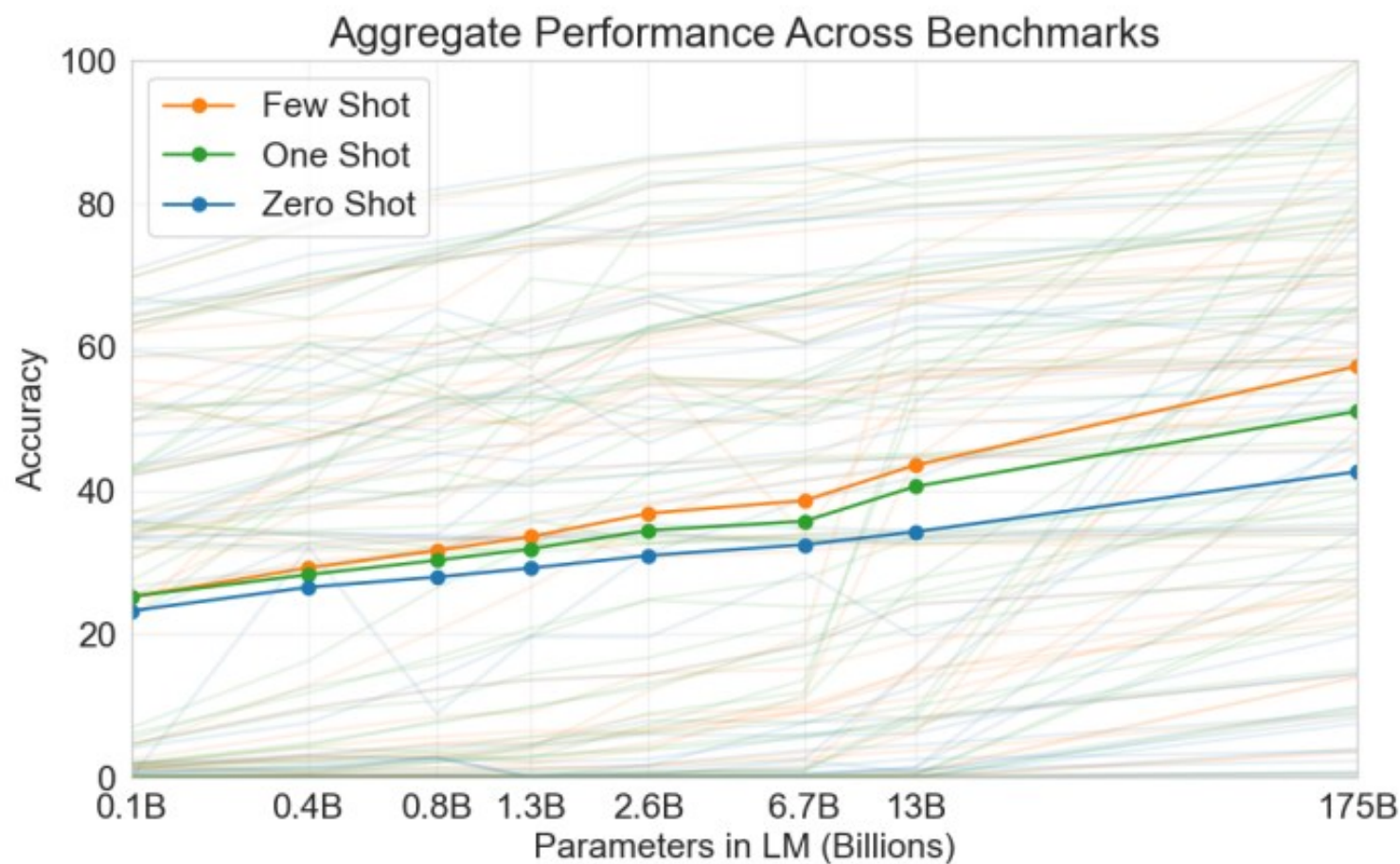


proteins

✳️ A stunning emergent property of Large Models ✳️

👉 Zero-shot/few-shot learning 👈

Large language models can tackle new tasks simply from natural language instructions



from Brown et al <https://arxiv.org/abs/2005.14165>

From Kaplan et al <https://arxiv.org/abs/2001.08361>

Larger models score **higher**, generalize **better**, train **faster**

The Lighton logo, consisting of the word 'Lighton' in white text on a black rectangular background, with a blue starburst icon replacing the letter 'o'.

March 2021: PAGnol, the largest model in French (1.5 billion parameters)

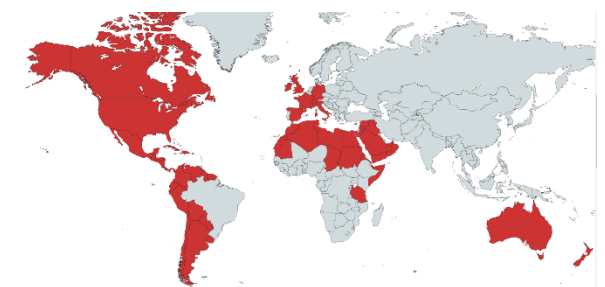


April 2022: Muse API and VLM-4 the **largest language models** in French, Italian, Spanish, German and Arabic



&

VLM4



👉 Zero-shot/few-shot learning 👈

 muse

&

VLM4


Demo Time! ✨

Completing Proust's "À l'ombre des jeunes filles en fleurs », in French



En rentrant, Françoise, me fit arrêter, au coin de la rue Royale, devant un étalage en plein vent où elle choisit, pour ses propres étrennes, des photographies de Pie IX et de Raspail, et où, pour ma part, j'en achetai une de la Berma. Les innombrables admirations qu'excitait l'artiste donnaient quelque chose d'un peu pauvre à ce visage unique qu'elle avait pour y répondre, immuable et précaire comme ce vêtement des personnes qui n'en ont pas de rechange, et où elle ne pouvait exhiber toujours que le petit pli au-dessus de la lèvre supérieure, le relèvement des sourcils, quelques autres particularités physiques toujours les mêmes qui, en somme, étaient à la merci d'une brûlure ou d'un choc. Ce visage

FR French

 Shuffle


Create

Completing Proust's "À l'ombre des jeunes filles en fleurs », in French



En rentrant, Françoise, me fit arrêter, au coin de la rue Royale, devant un étalage en plein vent où elle choisit, pour ses propres étrennes, des photographies de Pie IX et de Raspail, et où, pour ma part, j'en achetai une de la Berma. Les innombrables admirations qu'excitait l'artiste donnaient quelque chose d'un peu pauvre à ce visage unique qu'elle avait pour y répondre, immuable et précaire comme ce vêtement des personnes qui n'en ont pas de rechange, et où elle ne pouvait exhiber toujours que le petit pli au-dessus de la lèvre supérieure, le relèvement des sourcils, quelques autres particularités physiques toujours les mêmes qui, en somme, étaient à la merci d'une brûlure ou d'un choc. Ce visage par-dessus tout charmant était, pour moi, le signe qu'il m'était possible de connaître à peu près intimement cette personne dont toute une année j'avais cherché en vain le nom et les traits, et dont je commençais à croire qu'elle n'existait que dans ma pensée, et par chance, qui était aussi la plus gentille et la plus gaie des femmes. Je pris le tramway jusqu'à la rue de Richelieu, en face de la rue Cambon, et allai dîner au Lutetia. Je revins le lendemain, mais non sans un réel désir de revoir cette artiste, avec l'idée que je pourrais me risquer à bavarder un peu avec elle, à la regarder de près et sans lui adresser de mots d'amour. Il me fut impossible

FR French

 Shuffle

Create

Asking to classify a review after seeing a few examples, in **Italian**



Classifica le seguenti recensioni:

Recensione : Un libro straordinario, con una trama appassionante e ricca di colpi di scena

Giudizio : Positivo.

Recensione : Il terzo capitolo di questa saga è il culmine di una grande avventura. La storia è coinvolgente dall'inizio alla fine con un finale pieno di sorprese. Consiglio vivamente la lettura!

Giudizio : Positivo.

Recensione : Un romanzo estremamente noioso. Non ci sono personaggi interessanti, e la trama è banale e prevedibile

Giudizio : Negativo.

Recensione : Un libro terribile che non raccomanderei a nessuno

Giudizio :

IT Italian



Shuffle

Create

Asking to classify a review after seeing a few examples, in **Italian**



Classifica le seguenti recensioni:

Recensione : Un libro straordinario, con una trama appassionante e ricca di colpi di scena

Giudizio : Positivo.

Recensione : Il terzo capitolo di questa saga è il culmine di una grande avventura. La storia è coinvolgente dall'inizio alla fine con un finale pieno di sorprese. Consiglio vivamente la lettura!

Giudizio : Positivo.

Recensione : Un romanzo estremamente noioso. Non ci sono personaggi interessanti, e la trama è banale e prevedibile

Giudizio : Negativo.

Recensione : Un libro terribile che non raccomanderei a nessuno

Giudizio : Negativo.

IT Italian



Shuffle

Create

Finding the right keyword on a text about Picasso, in Spanish



Pablo Ruiz Picasso (Málaga, 25 de octubre de 1881–Mougins, 8 de abril de 1973) fue un pintor y escultor español, creador, junto con Georges Braque, del cubismo.

En el invierno de 1895 realizó su primer gran lienzo académico, La primera comunión (Museo Picasso, Barcelona), en Barcelona, ciudad en la que residió unos nueve años, salvo algunas vacaciones de verano y estancias más o menos largas en Madrid y París.

En enero de 1903 Picasso volvió a Barcelona. En primavera comenzó el cuadro La vida, uno de los mayores y más complejos lienzos de su época azul, considerado su trabajo más importante de estos años, obra de un simbolismo inusualmente oscuro en sus primeras obras y sujeto a múltiples interpretaciones académicas, sobre las cuales el artista nunca se pronunció.


Palabras clave:

- Picasso

-

es Spanish



 Shuffle

Create

Finding the right keyword on a text about Picasso, in Spanish



Pablo Ruiz Picasso (Málaga, 25 de octubre de 1881–Mougins, 8 de abril de 1973) fue un pintor y escultor español, creador, junto con Georges Braque, del cubismo.

En el invierno de 1895 realizó su primer gran lienzo académico, La primera comunión (Museo Picasso, Barcelona), en Barcelona, ciudad en la que residió unos nueve años, salvo algunas vacaciones de verano y estancias más o menos largas en Madrid y París.

En enero de 1903 Picasso volvió a Barcelona. En primavera comenzó el cuadro La vida, uno de los mayores y más complejos lienzos de su época azul, considerado su trabajo más importante de estos años, obra de un simbolismo inusualmente oscuro en sus primeras obras y sujeto a múltiples interpretaciones académicas, sobre las cuales el artista nunca se pronunció.

Palabras clave:

- Picasso
- Cubismo - Pintura - El Guernica - La vida

es Spanish



Shuffle

Create

Answering history questions, in German



Choose a language, shuffle for an example or produce your own, and enjoy the result.

Wer war Konrad Adenauer?

DE German



Shuffle

Create

Answering history questions, in German



Choose a language, shuffle for an example or produce your own, and enjoy the result.

Wer war Konrad Adenauer? Konrad Adenauer war der erste Bundeskanzler der Bundesrepublik Deutschland. Er wurde am 20. Januar 1876 in Köln geboren. Er war das vierte Kind von Adolf und Helene Adenauer. Adolf heiratete am 20. Januar 1895 in Köln die aus einer wohlhabenden Familie stammende Helene.

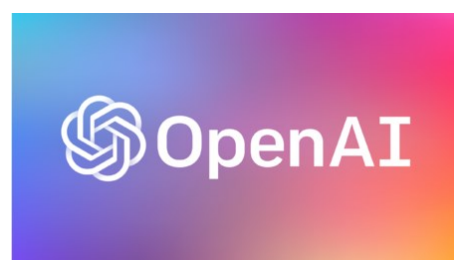
DE German



Shuffle

Create

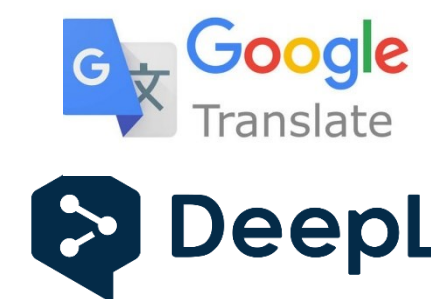
What about using GPT-3 in English and then simply translate into **French**, **Italian**,
Spanish or **German** ?



GPT3

&

Translation



instead of

Lighton **VLM4**



Quality comes in many shapes: local **knowledge/speak**

📍 Is the **Mont Saint Michel** in Normandy or Brittany? A (very) french **squabble...**



GPT-3
English, OpenAI

~~Brittany~~

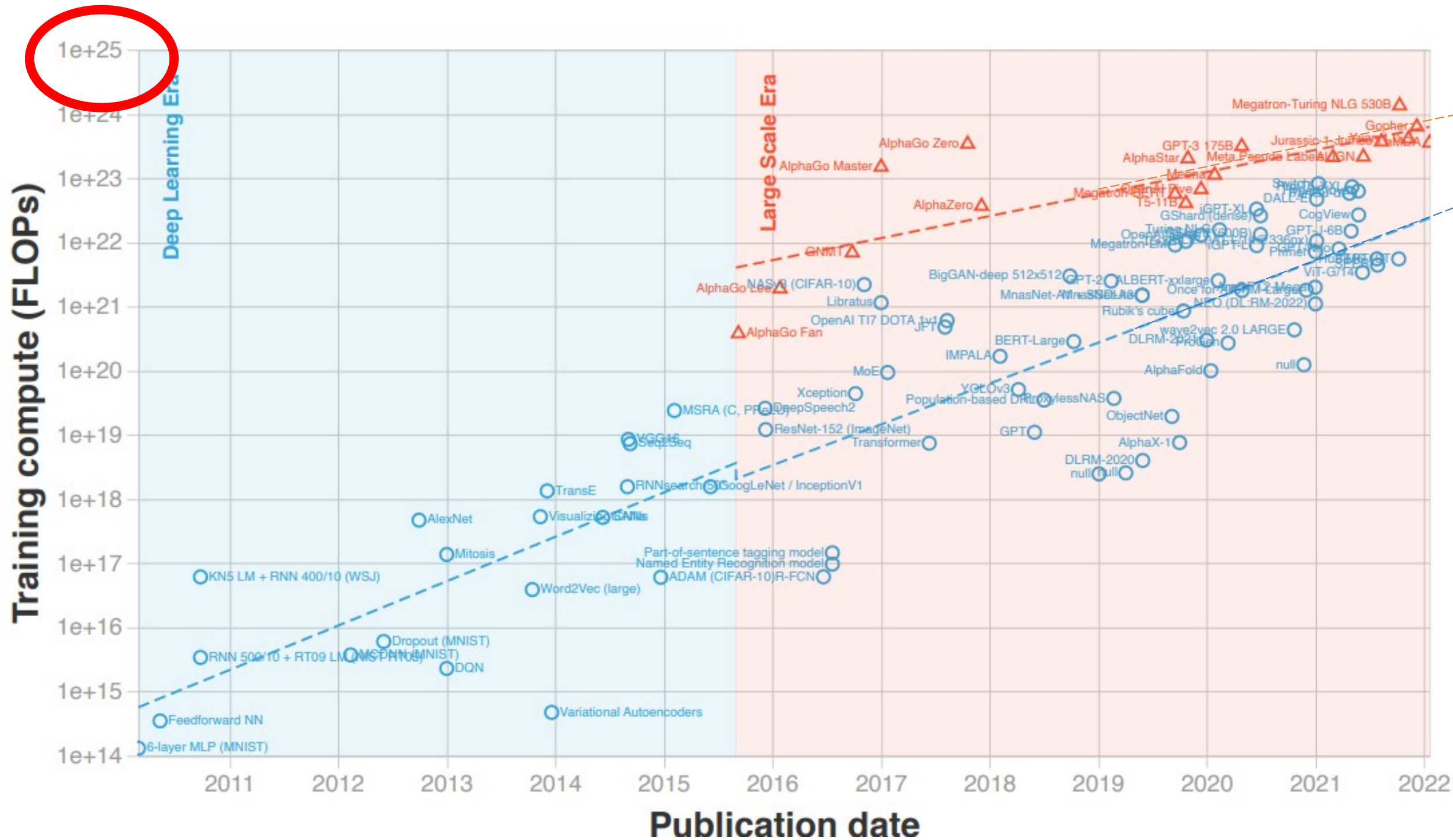
VLM-4
native French,
in Muse API

Normandy

👉 Also important for **local language variants**: **Januar** vs **Jänner**
“high-German” Austria

Limitation #1: Compute

10 000 000 000 000 000 000 000 000 000 000



cnrs INSTITUT DU DÉVELOPPEMENT ET DES RESSOURCES EN INFORMATIQUE SCIENTIFIQUE



EuroHPC Joint Undertaking



Limitation #2: Data

Do we have **enough data** to go beyond English? **Yes!**

Required **minimum** data

Model size	Minimum tokens
1.5B	20B
6.7B	37B
20B	55B
100B	100B
500B	181B

(Neural Scaling Laws, Kaplan et al.)

Data available in one year of **CommonCrawl**

Ranking	Language	High quality	Medium quality
1st	English	2T	6T
7th	French	260B	880B
15th	Indonesian	50B	145B
30th	Hindi	8B	16B

(1GB ~ 4B tokens, high quality top 20%, medium 20-40%, one dump every 3-4 months.)

from Launay, GTC'22

Limitation #3: Data Quality 📊

Quality comes in many shapes: maximising **zero-shot performance**

🌸 Example from Big Science:

BS-tr1-13B vs EAI-GPTNeo

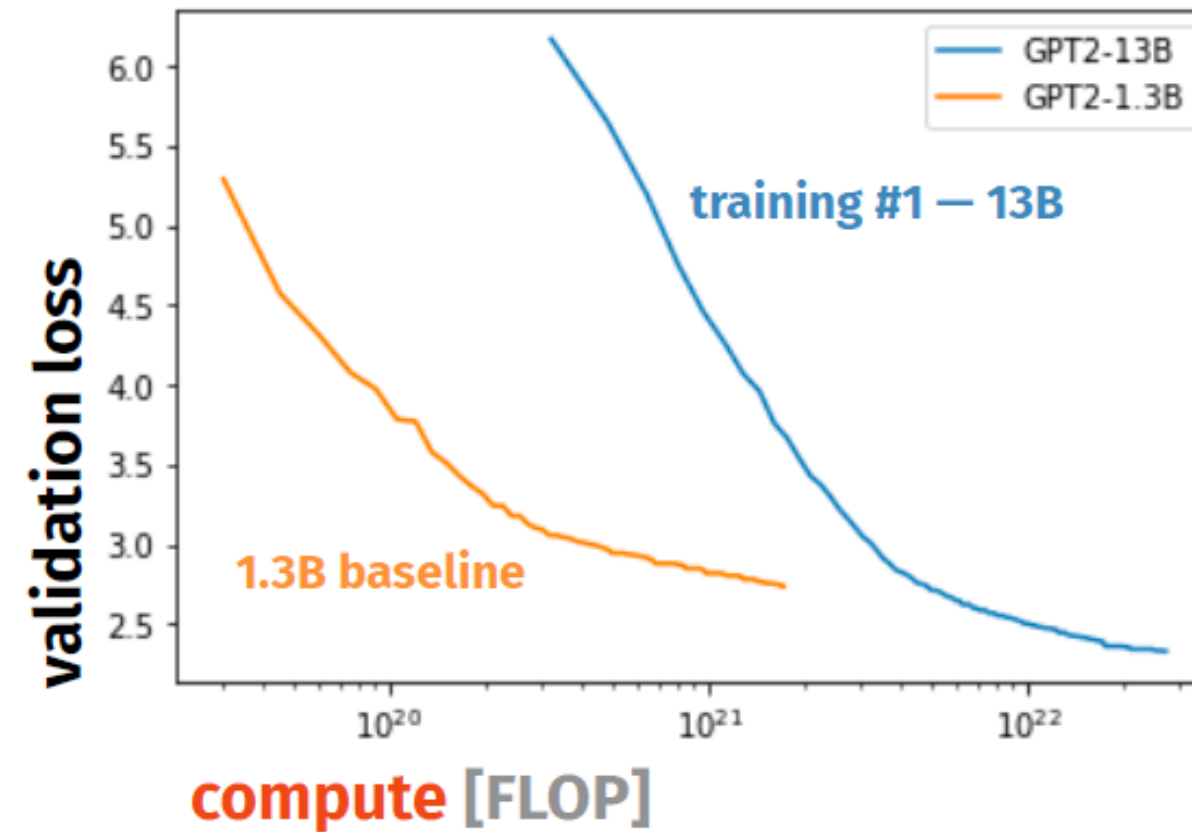
5.02	65%	54%
lambada ppl.	winogrande acc.	hellaswag acc.

more in line with a **1.3B-2.7B** model!

from debugging with 1.3B models...

The Pile >> OSCAR & C4

curated



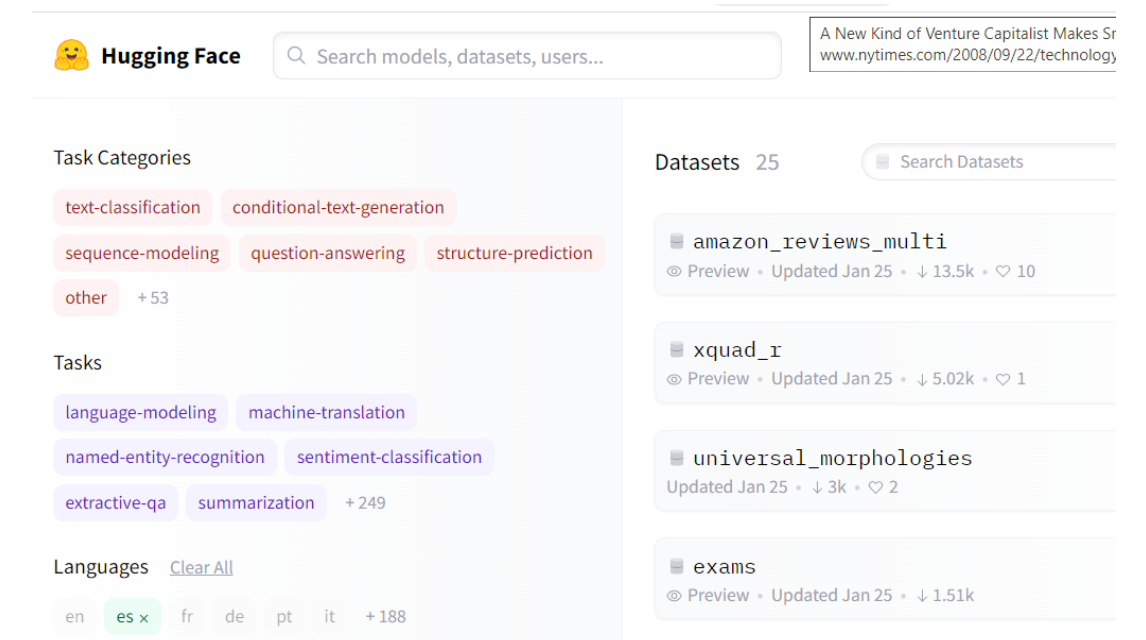
➡ Beyond English, large curated datasets are still **rare!**

from Launay, GTC'22

Limitation #4: Public high quality datasets for downstream task evaluation

English: 363 datasets

Spanish:	25	
Polish:	21	
German:	20	
Italian:	14	
French:	12	
Dutch:	5	
Czech:	3	
Slovenian:	1	



Quality: 12 datasets in French are of higher quality than 14 in Italian !!

Limitation #5: Critical mass for certain languages and markets ?

OpenAI

ANTHROPIC

On the Opportunities and Risks of Foundation Models

Rishi Bommasani* Drew A. Hudson Ehsan Adeli Russ Altman Simran Arora
 Sydney von Arx Michael S. Bernstein Jeannette Bohg Antoine Bosselut Emma Brunskill
 Erik Brynjolfsson Shyamal Buch Dallas Card Rodrigo Castellon Niladri Chatterji
 Annie Chen Kathleen Creel Jared Quincy Davis Dorothea Demszky Chris Donahue
 Moussa Doumbouya Esin Durmus Stefano Ermon John Etchemendy Kawin Ethayarajh
 Li Fei-Fei Chelsea Finn Trevor Gale Lauren Gillespie Karan Goel Noah Goodman
 Shelby Grossman Neel Guha Tatsunori Hashimoto Peter Henderson John Hewitt
 Daniel E. Ho Jenny Hong Kyle Hsu Jing Huang Thomas Icard Saahil Jain
 Dan Jurafsky Pratyusha Kalluri Siddharth Karamcheti Geoff Keeling Fereshte Khani
 Omar Khattab Pang Wei Koh Mark Krass Ranjay Krishna Rohith Kudithipudi
 Ananya Kumar Faisal Ladhak Mina Lee Tony Lee Jure Leskovec Isabelle Levent
 Xiang Lisa Li Xuechen Li Tengyu Ma Ali Malik Christopher D. Manning
 Suvir Mirchandani Eric Mitchell Zanele Munyikwa Suraj Nair Avanika Narayan
 Deepak Narayanan Ben Newman Allen Nie Juan Carlos Niebles Hamed Nilforoshan
 Julian Nyarko Giray Ogut Laurel Orr Isabel Papadimitriou Joon Sung Park Chris Piech
 Eva Portelance Christopher Potts Aditi Raghunathan Rob Reich Hongyu Ren
 Frieda Rong Yusuf Roohani Camilo Ruiz Jack Ryan Christopher Ré Dorsa Sadigh
 Shiori Sagawa Keshav Santhanam Andy Shih Krishnan Srinivasan Alex Tamkin
 Rohan Taori Armin W. Thomas Florian Tramèr Rose E. Wang William Wang Bohan Wu
 Jiajun Wu Yuhuai Wu Sang Michael Xie Michihiro Yasunaga Jiaxuan You Matei Zaharia
 Michael Zhang Tianyi Zhang Xikun Zhang Yuhui Zhang Lucia Zheng Kaitlyn Zhou
 Percy Liang*¹

Center for Research on Foundation Models (CRFM)
 Stanford Institute for Human-Centered Artificial Intelligence (HAI)
 Stanford University



A Roadmap for Big Model *

Sha Yuan^{*1} Hanyu Zhao^{*1} Shuai Zhao^{*1} Jiahong Leng^{*1} Yangxiao Liang^{*1} Xiaozhi Wang^{*2} Jifan Yu^{*2} Xin Lv^{*2}
 Zhou Shao^{*1} Jiaao He^{*2} Yankai Lin^{*3} Xu Han^{*2} Zhenghao Liu^{*4} Ning Ding^{*2} Yongning Rao^{*2} Yizhao Gao^{*5}
 Liang Zhang^{*5} Ming Ding^{*2} Cong Fang^{*6} Yisen Wang^{*6} Mingsheng Long^{*2} Jing Zhang^{*5} Yinpeng Dong^{*2} Tianyu
 Pang^{*2} Peng Cui^{*2} Lingxiao Huang^{*7} Zheng Liang^{*2} Huawei Shen^{*8} Hui Zhang^{*2} Quanshi Zhang^{*9} Qingxiu Dong^{*6}
 Zhixing Tan^{*2} Mingxuan Wang^{*13} Shuo Wang^{*2} Long Zhou^{*14} Haoran Li^{*10} Junwei Bao^{*10} Yingwei Pan^{*10} Weinan
 Zhang^{*11} Zhou Yu^{*12} Rui Yan^{*5} Chence Shi^{*15} Minghao Xu^{*15} Zuo Bai Zhang^{*15} Guoqiang Wang^{*1} Xiang Pan^{*16}
 Mengjie Li^{*17} Xiaoyu Chu^{*1} Zijun Yao^{*2} Fangwei Zhu^{*2} Shulin Cao^{*2} Weicheng Xue^{*2} Zixuan Ma^{*2} Zhengyan Zhang^{*2}
 Shengding Hu^{*2} Yujia Qin^{*2} Chaojun Xiao^{*2} Zheni Zeng^{*2} Ganqu Cui^{*2} Weize Chen^{*2} Weilin Zhao^{*2} Yuan Yao^{*2} Peng Li³
 Wenzhao Zheng^{*2} Wenliang Zhao^{*2} Ziyi Wang^{*2} Borui Zhang^{*2} Nanyi Fei^{*5} Anwen Hu^{*5} Zenan Ling^{*6} Haoyang Li^{*5} Boxi
 Cao^{*18} Xianpei Han^{*18} Weidong Zhan^{*6} Baobao Chang^{*6} Hao Sun^{*2} Jiawen Deng^{*2} Chujiu Zheng^{*2} Juanzi Li^{*22} Lei Hou^{*22}
 Xigang Cao^{*21} Jidong Zhu^{*22} Zhiyuan Liu^{*22} Maosong Sun^{*22} Jiwen Lu^{*22} Zhiwu Li^{*25} Qin Jin^{*25} Ruihua Song^{*5}
 Ji-Rong Wen^{*5} Zhouchen Lin^{*26} Liwei Wang^{*26} Hang Su^{*22} Jun Zhu^{*22} Zhifang Su^{*26} Jiajun Zhang^{*219} Yang Liu^{*22}
 Xiaodong He^{*210} Minlie Huang^{*22} Jian Tang^{*215} **Jie Tang**^{*22,1}

- ¹ Beijing Academy of Artificial Intelligence
- ² Tsinghua University
- ³ Wechat, Tencent Inc.
- ⁴ Northeastern University
- ⁵ Renmin University of China
- ⁶ Peking University
- ⁷ Huawei TCS Lab
- ⁸ Institute of Computing Technology, Chinese Academy of Sciences
- ⁹ Shanghai Jiao Tong University
- ¹⁰ JD AI Research
- ¹¹ Harbin Institute of Technology
- ¹² Columbia University
- ¹³ ByteDance AI Lab
- ¹⁴ Microsoft Research Asia
- ¹⁵ Mila-Quebec AI Institute & University of Montreal
- ¹⁶ New York University
- ¹⁷ BeiHang University
- ¹⁸ Institute of Software, Chinese Academy of Sciences
- ¹⁹ Institute of Automation Chinese Academy of Sciences

Jie Tang designs this big model roadmap. Authors labeled with ES organize different parts of this report. Authors labeled with * contribute equally. They are ranked according to their section.

AI21 labs

Large Language Models

- **New and seamless way of interacting with machines and the world around us.**
- **Already** make a difference how **we do Science** and
- **Soon** how we will **do businesss**, work and **learn**

With 20+ languages of varying level of representation, Europe can make a powerful move by **standardizing training and evaluation datasets for all**

- **strengthen** internal market
- **enhance competitiveness** of EU in the world

Thank you!

 Muse.LightOn.ai 

Igor Carron, PhD

Light^{on}, CEO

contact@lighton.ai