In the name of GOD

On the Opportunities and Risks of Foundation Models

Maryam Afshari

1 Introduction

Our Goals at a Glance

1

Emergence & homogenization

2

Social impact & foundation models ecosystem

3

The future of foundation models

4

Overview of this report


Microsoft
Turing-NLG


OpenAI
GPT-3, DALL-E, CLIP, Codex


AI21labs
Jurassic-1
ANTHROP\C


Google
BERT, T5, LaMDA, MUM


facebook
BART, RoBERTa, XLM


NAVER
HyperCLOVA
BAAI
Wu Dao 2.0

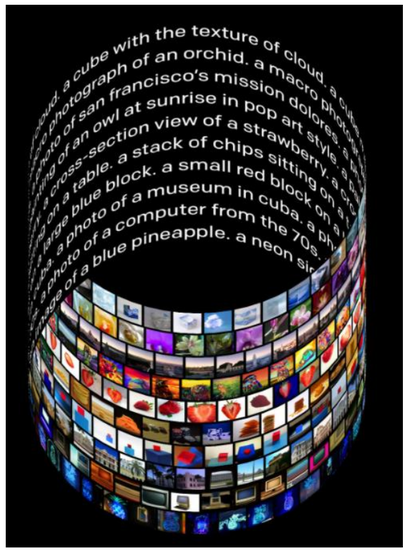

HUAWEI
PanGu-Alpha


NVIDIA
MegatronLM

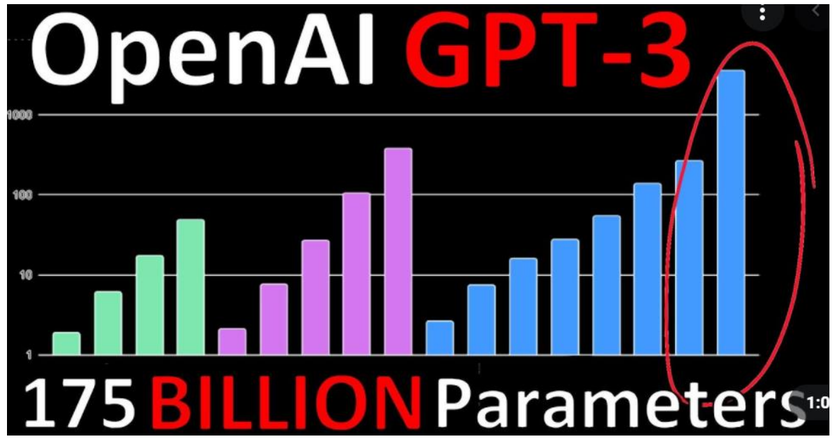
Foundation Models?

rise of models (e.g., BERT, DALL-E, GPT-3) that are trained on broad data at scale and are adaptable to a wide range of downstream tasks

DALL·E: Creating Images from Text



scale



Emergence

how a task is performed emerges (is inferred automatically) from examples -> ML

high-level features used for prediction emerge -> Deep learning

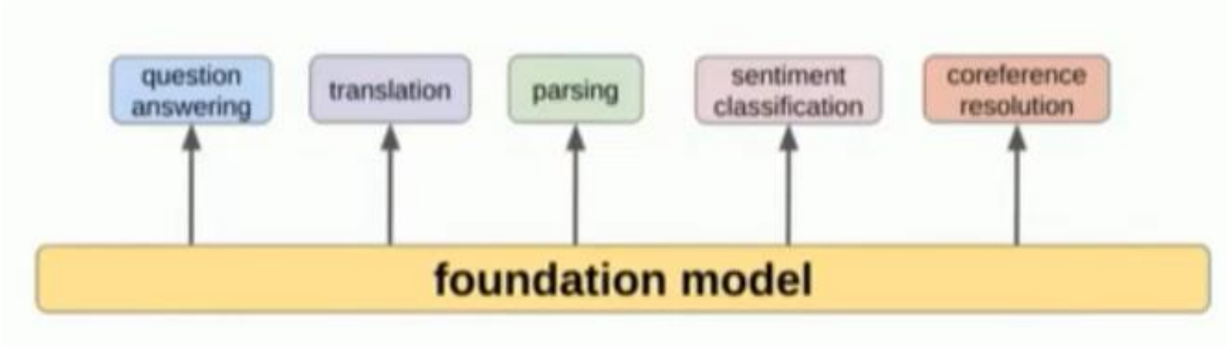
advanced functionalities such as in-context learning emerge-> Foundation models

Homogenization

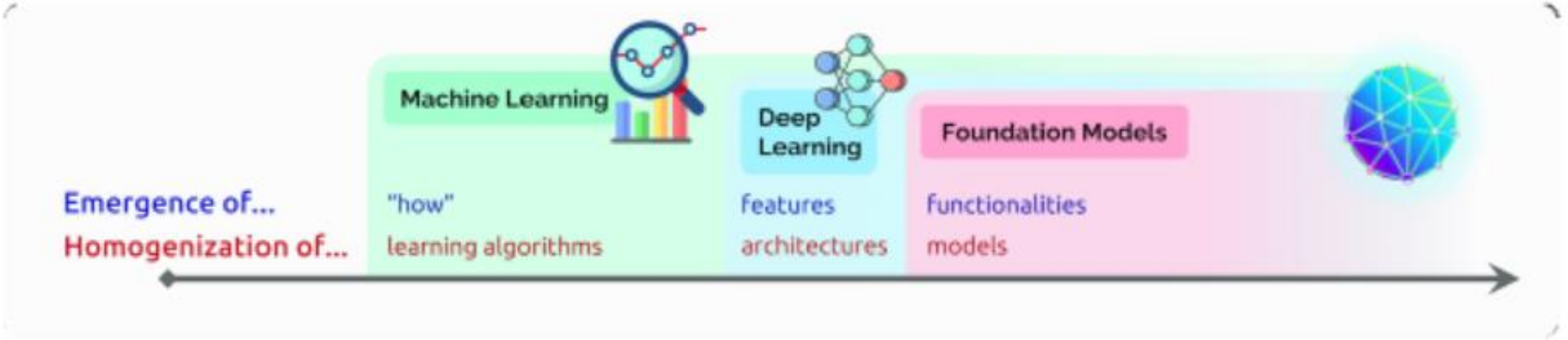
homogenizes learning algorithms (e.g., logistic regression)-> ML

homogenizes model architectures (e.g., Convolutional Neural Networks)-> Deep learning

foundation models homogenizes the model itself (e.g., GPT-3)> Foundation models



Emergence & homogenization



Transfer learning

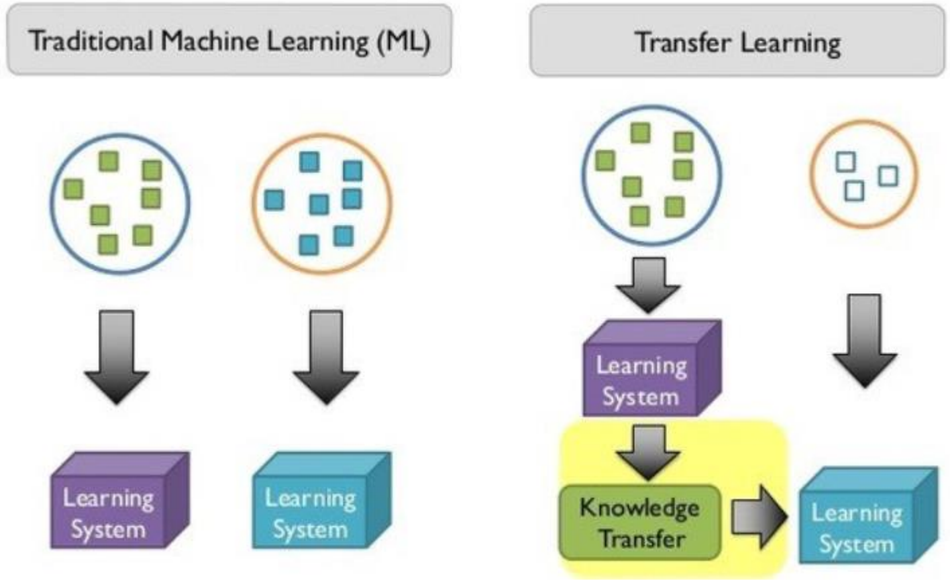
Foundation models have taken shape most strongly in NLP

By the end of 2018, the beginning of the era of foundation models

foundation models are enabled by transfer learning and scale

Transfer learning is what makes foundation models possible

scale is what makes them powerful



Pretrain & fine-tuning ← to take the “knowledge” learned from one task (e.g., object recognition in images) and apply it to another task (e.g., activity recognition in videos).

Scale

scale is what makes them powerful

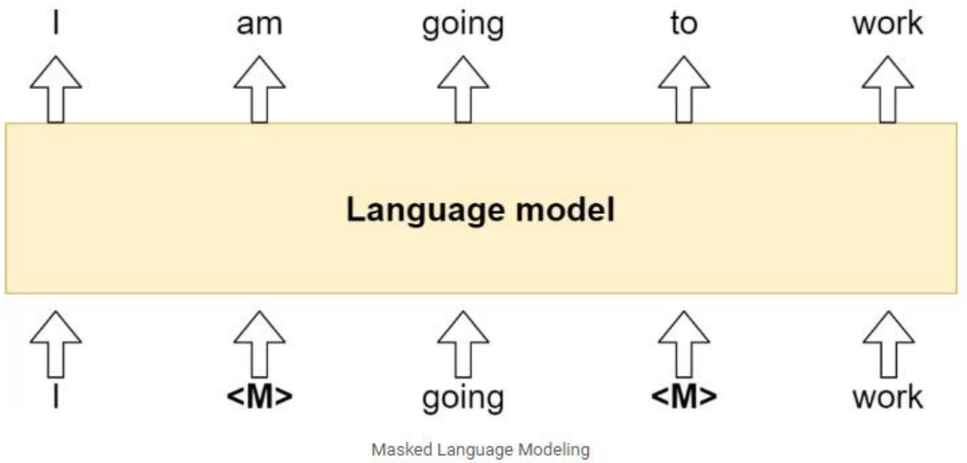
Scale required three ingredients:

1. **improvements in computer hardware — e.g., GPU throughput and memory have increased 10× over the last four years**

2. **the development of the Transformer model architecture that leverages the parallelism of the hardware to train much more expressive models than before**

3. **the availability of much more training data.**

self-supervised learning



Masked Language Modeling

In self-supervised learning on the other hand, the **pretraining task** is derived automatically from **unannotated data**

For example, the masked language modeling task used to train is to predict a missing word in a sentence given its surrounding context (e.g., I like_____sprouts)

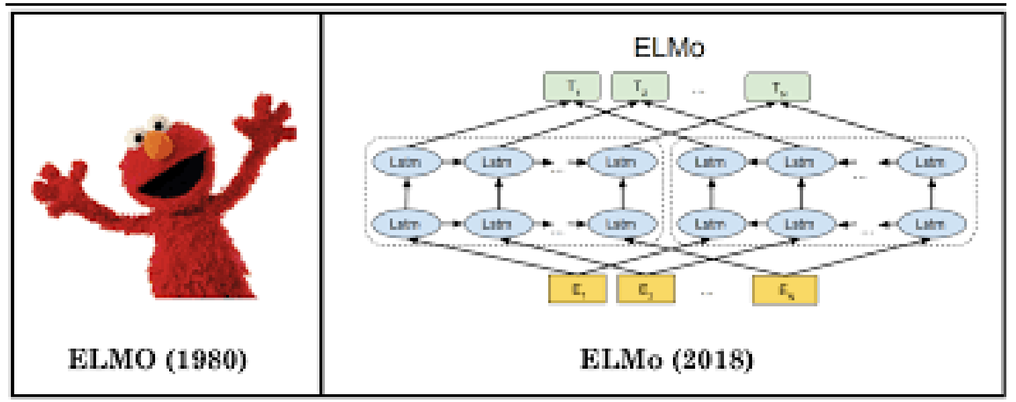
Self-supervised tasks are not only **more scalable**, only depending on **unlabeled data**, but they are designed to force the model **to predict parts of the inputs**, making them **richer and potentially more useful** than models trained on a more limited label space.

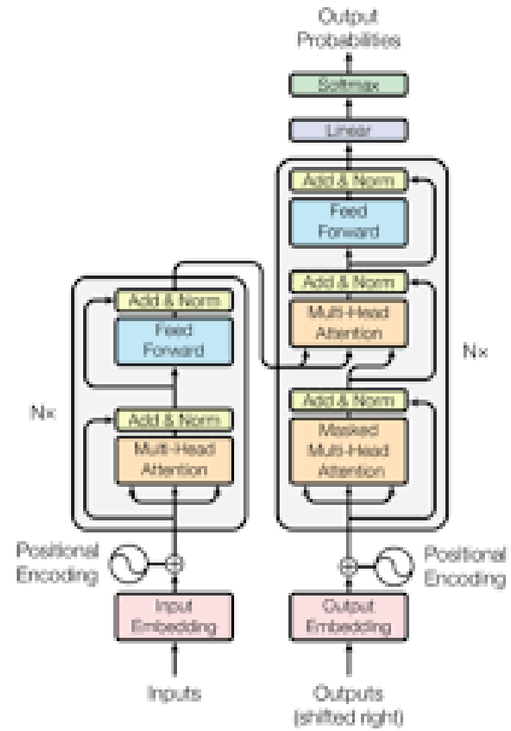
Word embedding

There had been considerable progress in self-supervised learning dating back to **word embeddings** which associated **each word with a context-independent vector**, provided the basis for a wide range of NLP models.

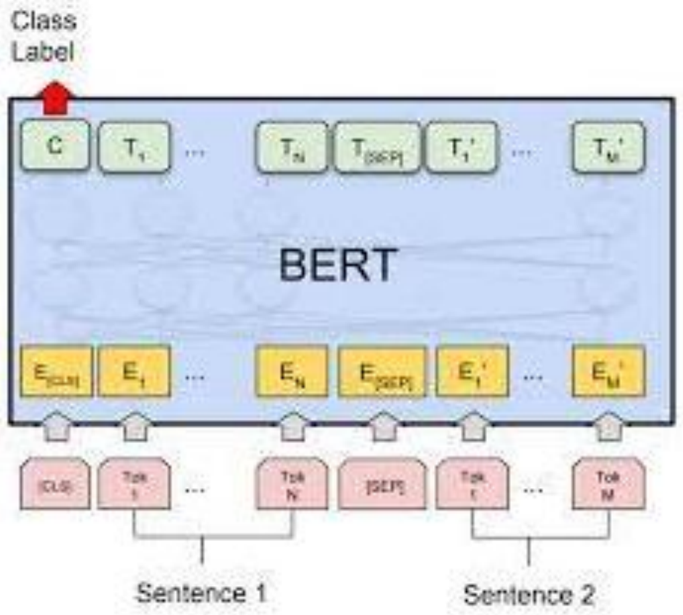
Shortly thereafter, self-supervised learning based on **autoregressive language modeling** (predict the next word given the previous words) became popular. This produced models that represented words in context, such as **GPT** , **ELMo** , and **ULMFiT** .

The **next wave** of developments **in self-supervised learning** — BERT GPT-2, RoBERTa , T5, BART —quickly followed, embracing the **Transformer architecture, incorporating** more powerful **deep bidirectional encoders** of sentences, and scaling up to larger models and datasets





BERT



While one **can view this last wave** of technical developments purely through the **lens of selfsupervised learning**, there was **a sociological inflection** point around the **introduction of BERT.**

Before 2019, **self-supervised learning with language models** was essentially a subarea in NLP, which progressed in parallel to other developments in NLP.

After 2019, self-supervised learning with language models became more of a **substrate of NLP,** as using BERT has become the norm.

The acceptance **that a single model could be useful for such a wide range of tasks** marks the beginning of the **era of foundation models**.

Foundation models (liability)

Foundation models have led to an **unprecedented level of homogenization**: Almost all **state-of the-art NLP models** are now adapted from one of **a few foundation models,** such as BERT, RoBERTa,BART, T5, etc.

While this homogenization **produces extremely high leverage** (any improvements in the foundation models can lead to immediate benefits across all of NLP), it is also **a liability**; all AI systems might **inherit** the same problematic biases of a few foundation models.

Foundation models (research communities)

We are also beginning to see a homogenization across research communities. For example, similar **Transformer-based sequence modeling approaches** are now applied to **text** [Devlin et al. 2019;Radford et al. 2019; Raffel et al. 2019], **images** [Dosovitskiy et al. 2020; Chen et al. 2020d], **speech** [Liu et al. 2020d], tabular data [Yin et al. 2020], protein sequences [Rives et al. 2021], organic molecules[Rothchild et al. 2021], and reinforcement learning [Chen et al. 2021b; Janner et al. 2021].

These examples point to a **possible future** where we have **a unified set of tools** for **developing foundation models** across a wide range of **modalities** [Tamkin et al. 2021a]

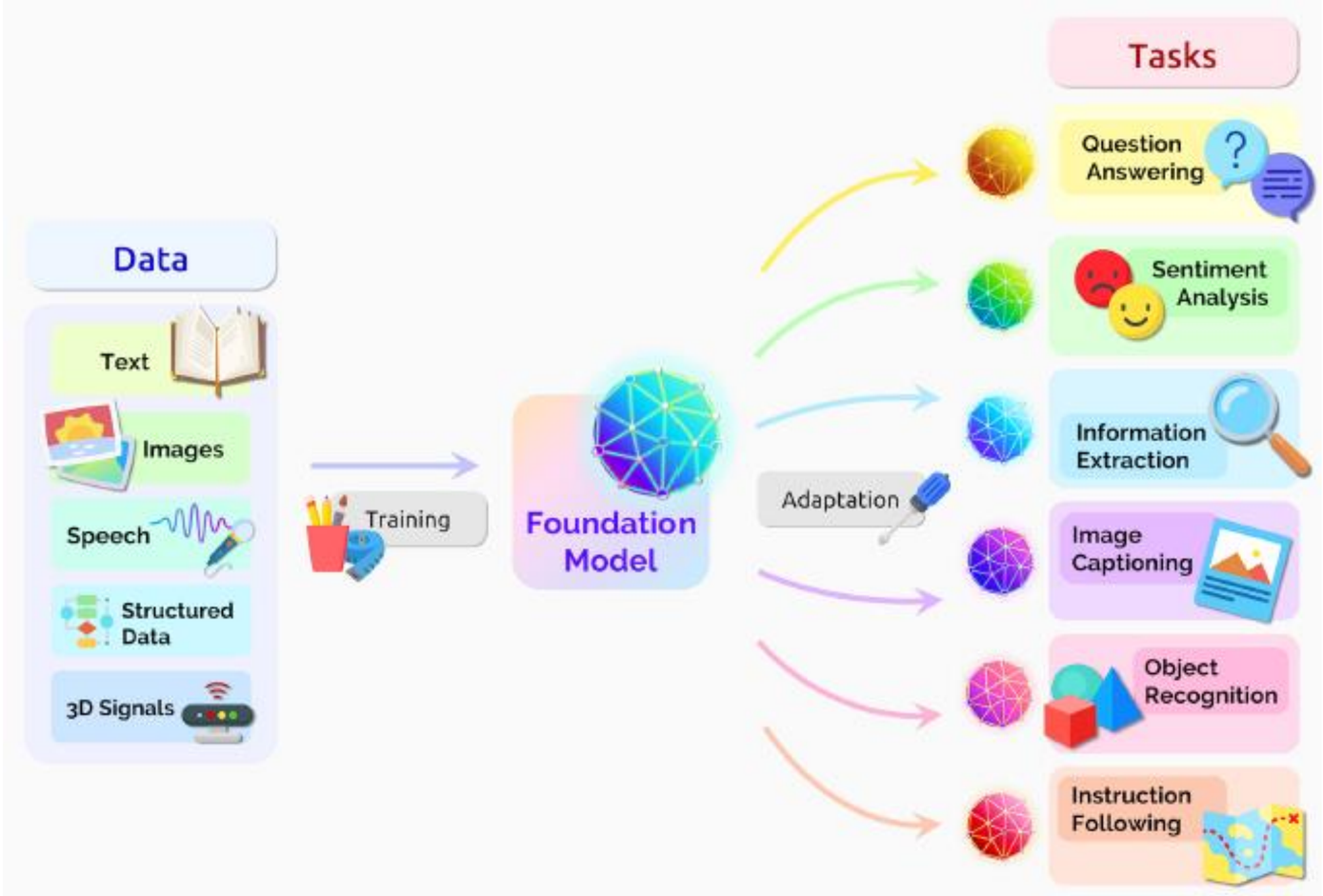
Besides **the homogenization of approaches**, we also see the **homogenization of actual models** across research communities in the form of **multimodal models** — e.g., **foundation models trained on language and vision data** [Luo et al. 2020; Kim et al. 2021a; Cho et al. 2021; Ramesh et al. 2021; Radford et al. 2021].

multimodal Foundation models

Data is naturally multimodal in some domains—e.g., medical images, structured data, clinical text in healthcare (§3.1: healthcare).

Thus, **multimodal foundation models** are a natural way of **fusing all the relevant information** about a domain, and adapting to tasks that also span multiple modes.

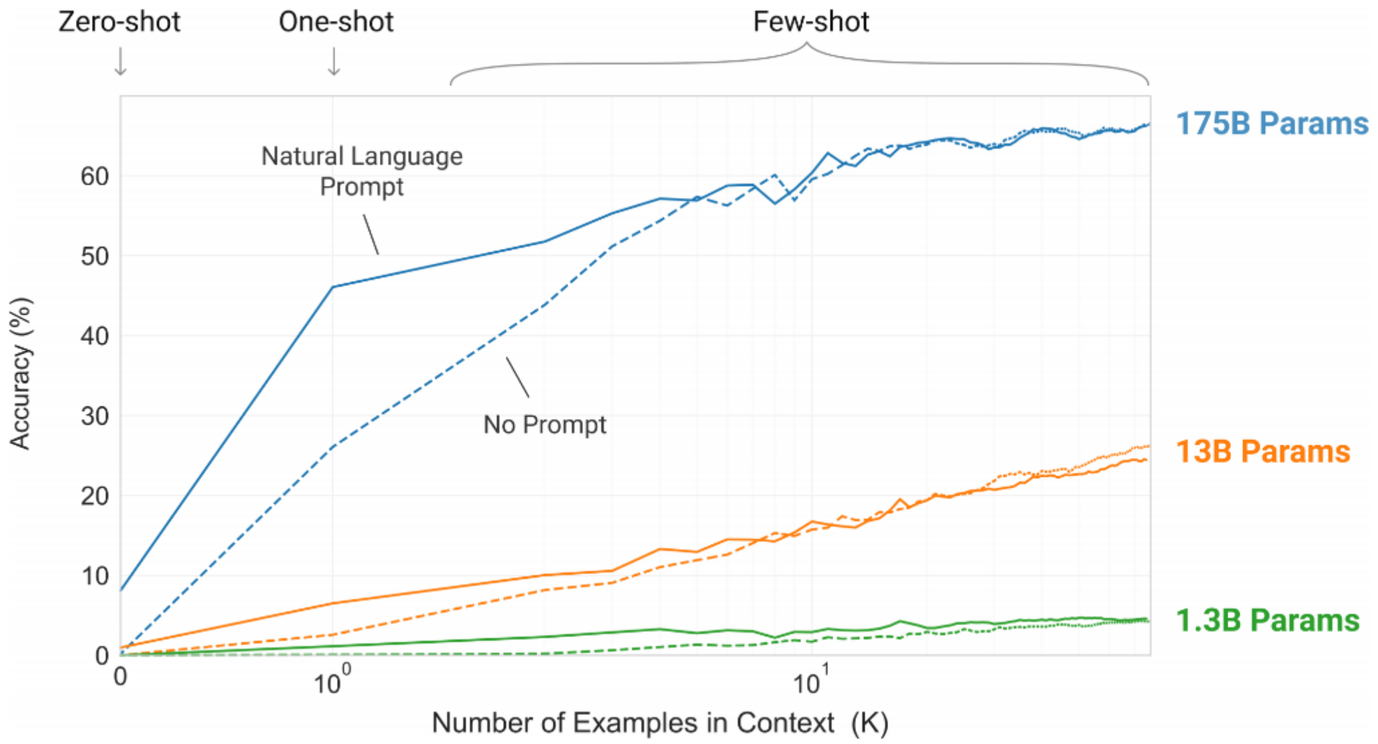
A foundation model can **centralize** the information from **all the data from various modalities**. This one model can then be adapted to a **wide range of downstream tasks**.



emergence in foundation models

Foundation models have also led to **surprising emergence** which results from **scale**.

For example,**GPT-3** , with **175 billion** parameters compared to **GPT-2**'s **1.5 billion**, permits **in-context learning**, in which the language model can be adapted to a downstream task simply by providing it with **a prompt (a natural language description of the task**), **an emergent property** that was neither specifically trained for nor anticipated to arise.



Interaction (Homogenization and emergence)

Homogenization and emergence interact in a potentially unsettling way.

Homogenization could potentially provide enormous gains for many domains where task-specific data is quite limited see the opportunities presented in several such domains (e.g., §3.1: healthcare, §3.2: law, §3.3: education);

on the other hand, any **flaws in the model** are blindly inherited by all adapted models (§5.1: fairness, §5.6: ethics).

Since the **power of foundation models** comes from their **emergent qualities** rather than their explicit construction, existing **foundation models are hard to understand** (§4.4: evaluation, §4.10: theory, §4.11: interpretability)
and they have **unexpected failure modes** (§4.7: security, §4.8: robustness).

Since **emergence** generates substantial u**ncertainty over the capabilities** and **flaws of foundation** models, **aggressive homogenization** through these models is risky business.

Derisking is the central challenge in the further development of foundation models from an ethical (§5.6: ethics) and AI safety (§4.9: ai-safety) perspective.

1.1

Naming

- **foundation models**
- *Pre-trained model, self-supervised model*
- *Language model*
- *general-purpose model and multi-purpose model*
- *task-agnostic model*

- partially **capture the technical dimension** of these models, **but fail to** capture the significance of the paradigm shift in an **accessible manner for those beyond machine learning**.
- is too **narrow**; as we describe, the scope of foundation models goes well beyond language.
- that capture the **important aspect** that these models can serve **multiple downstream tasks**, but both fail to capture their **unfinished character** and **the need for adaptation**.
- model would capture **the manner of training**, but fail to capture the **significant implication to downstream applications**.

1.1

Naming

• foundation models

- We chose the new term foundation models to **identify the models and the emerging paradigm** that are the subject of this report.
- In particular, the word “**foundation**” specifies ***the role these models play***: a foundation model is itself incomplete but serves as the common basis from which many task-specific models are built via adaptation.
- We also chose the term “foundation" to connote the significance of **architectural stability, safety, and security**: poorly-constructed foundations are a recipe for disaster and well-executed foundations are a reliable bedrock for future applications.
- At present, we emphasize that **we do not fully understand the nature or quality of the foundation** that foundation models provide;
- we cannot characterize whether the foundation is trustworthy or not.
- Thus, **this is a critical problem for researchers**, foundation model providers, application developers who rely on foundation models, policymakers, and society at large to address.

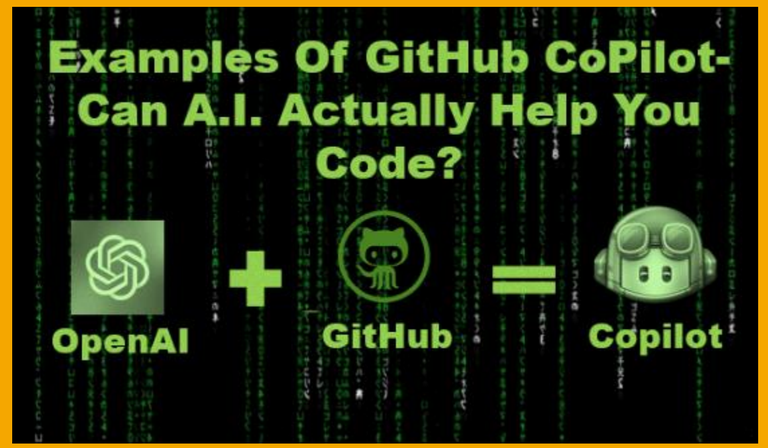
2 Social impact and the foundation models ecosystem

Development vs. Deployment

Research models
are often not extensively tested and might have unknown failure modes; warning labels should be placed on research models that are not fit to deploy.

deployed foundation
models that actually affect people's lives should be subject to much more rigorous testing and auditing.

To further understand the research and deployment of foundation models, we must zoom out and **consider the full ecosystem** that these foundation models inhabit, from data creation to actual deployment.



(Google search using BERT).

Social impact

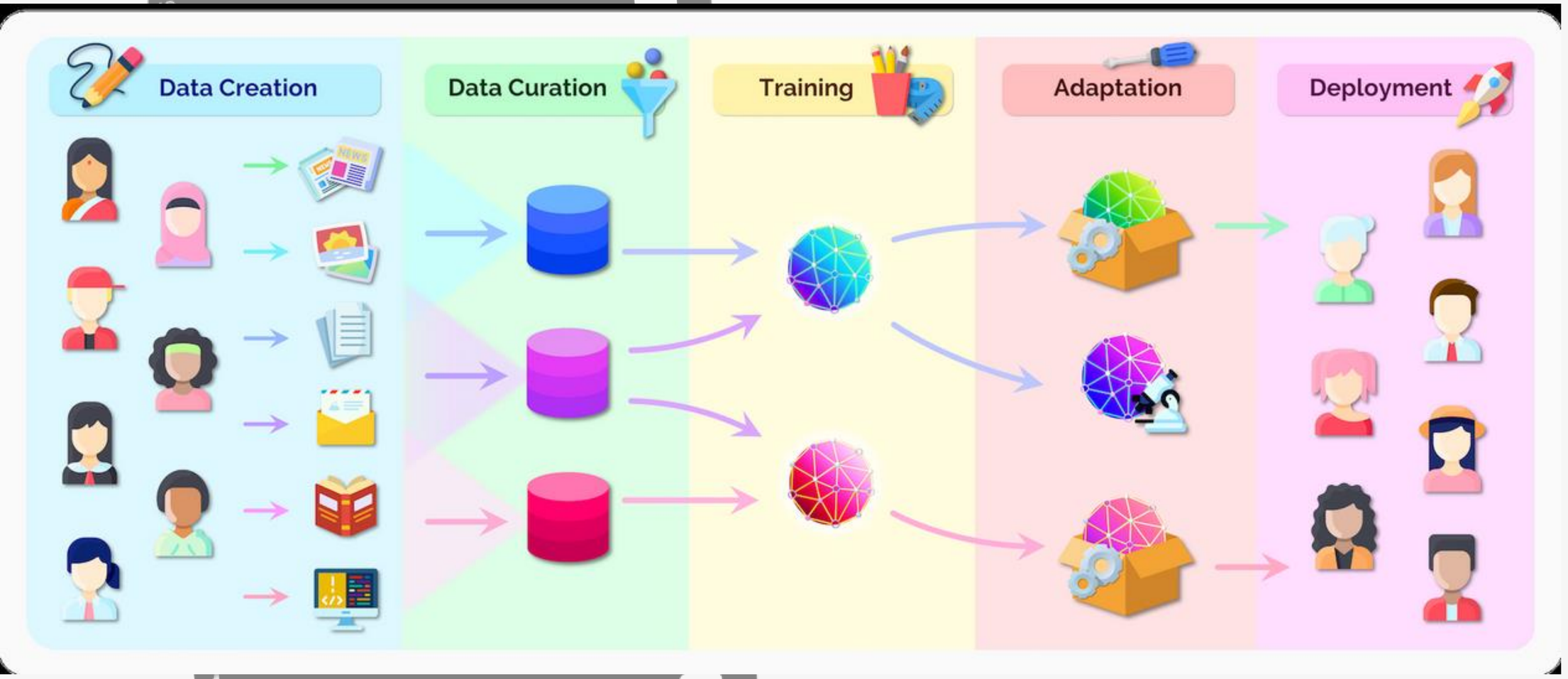
1 **Data creation**

2 **Data curation**

3 **Training**

4 **Adaptation** **(e.g., document summarization).**

5 **Deployment**


Data Creation
Data Curation
Training
Adaptation
Deployment

3

The future of foundation models

- **Disciplinary diversity**
- **Incentives**
- **Loss in accessibility**

loss in accessibility

deep learning: very open

foundation models: more closed

requires immense resources

Think ecosystem, act model

surrogate metrics for a representative set of potential downstream evaluation

a commitment to documenting these metrics

4 Overview of this report

The writing of this report was an experiment: we had over 100 people from different backgrounds **come together to write single report covering a wide range of aspects of foundation models.**

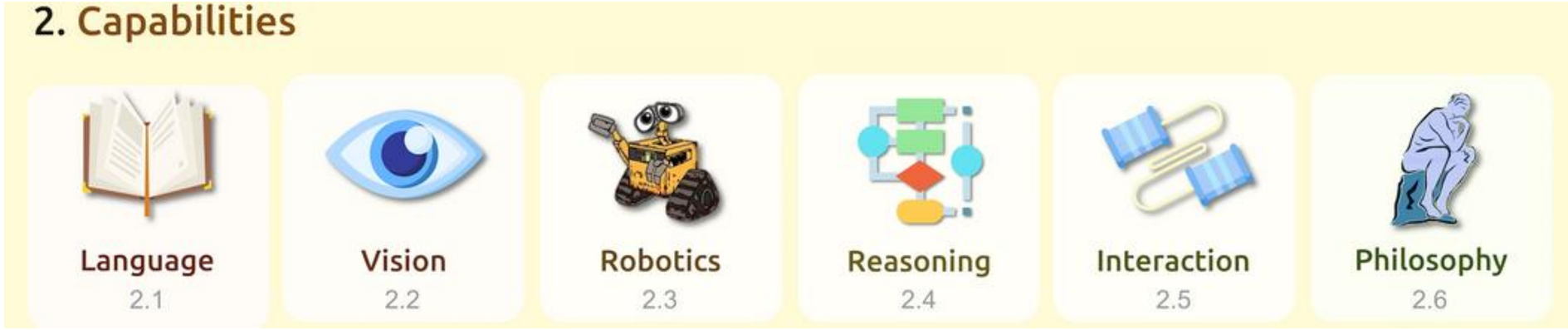
On the Opportunities and Risks of Foundation Models

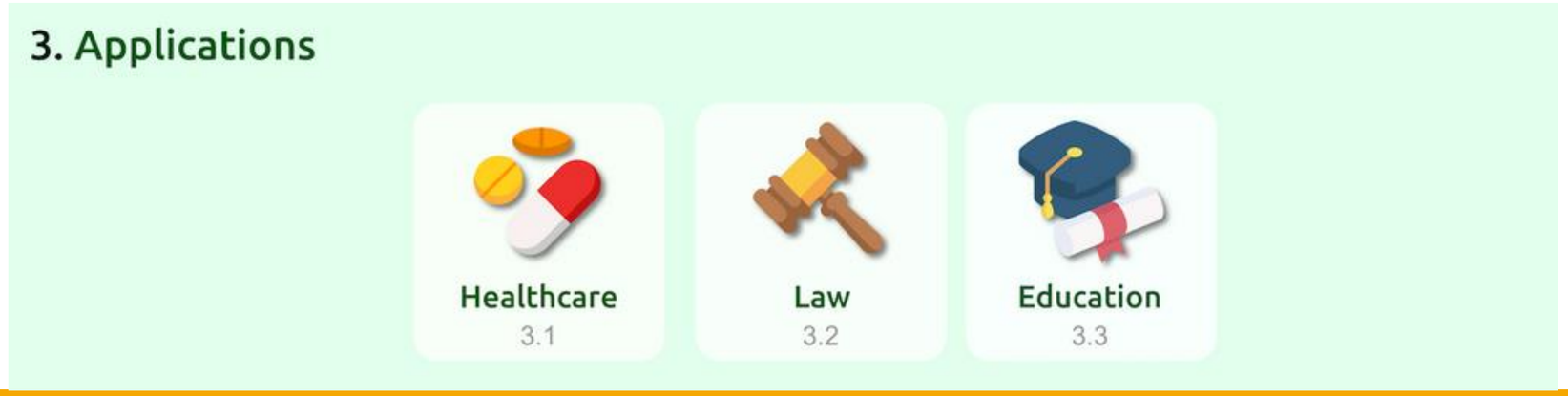
Rishi Bommasani* Drew A. Hudson Ehsan Adeli Russ Altman Simran Arora
Sydney von Arx Michael S. Bernstein Jeannette Bohg Antoine Bosselut Emma Brunskill
Erik Brynjolfsson Shyamal Buch Dallas Card Rodrigo Castellon Niladri Chatterji
Annie Chen Kathleen Creel Jared Quincy Davis Dorottya Demszky Chris Donahue
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 Kuditipudi
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

Structure

The report is divided **into 26 sections,** each discussing one aspect of foundation models. The section are grouped into four parts :






Thank you!