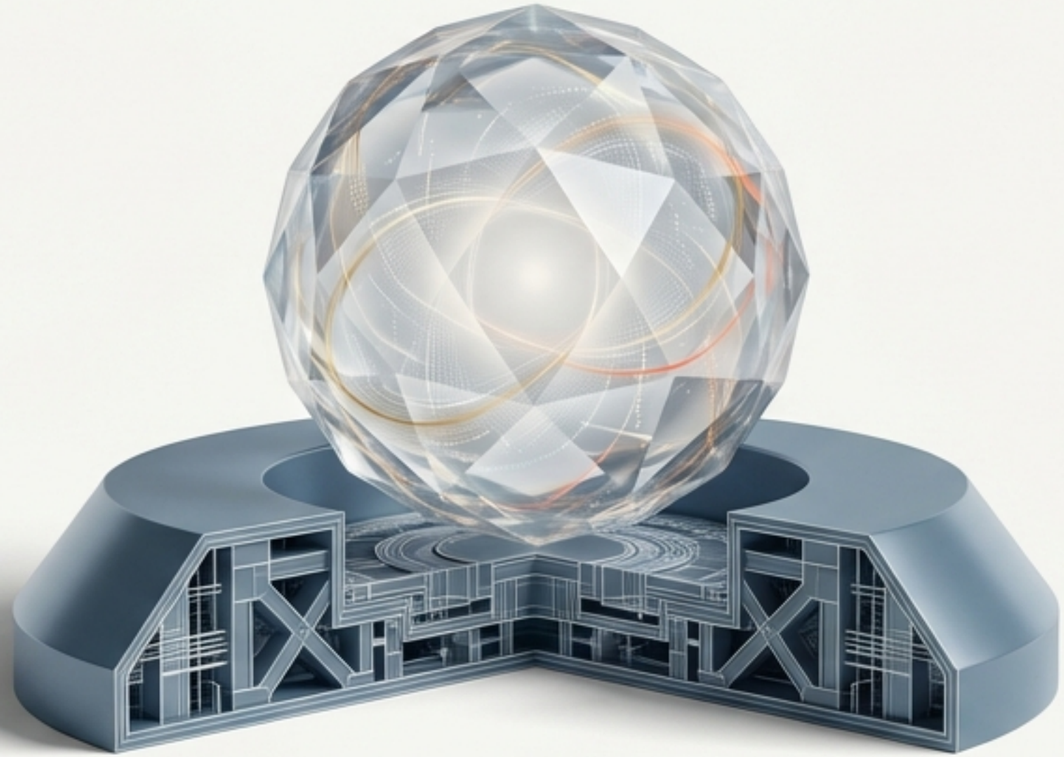
On the Opportunities and Risks of Foundation Models

A Report from the Center for Research on Foundation Models (CRFM)
Stanford Institute for Human-Centered Artificial Intelligence (HAI)
Rishi Bommasani, Percy Liang, et al. (over 100 authors)

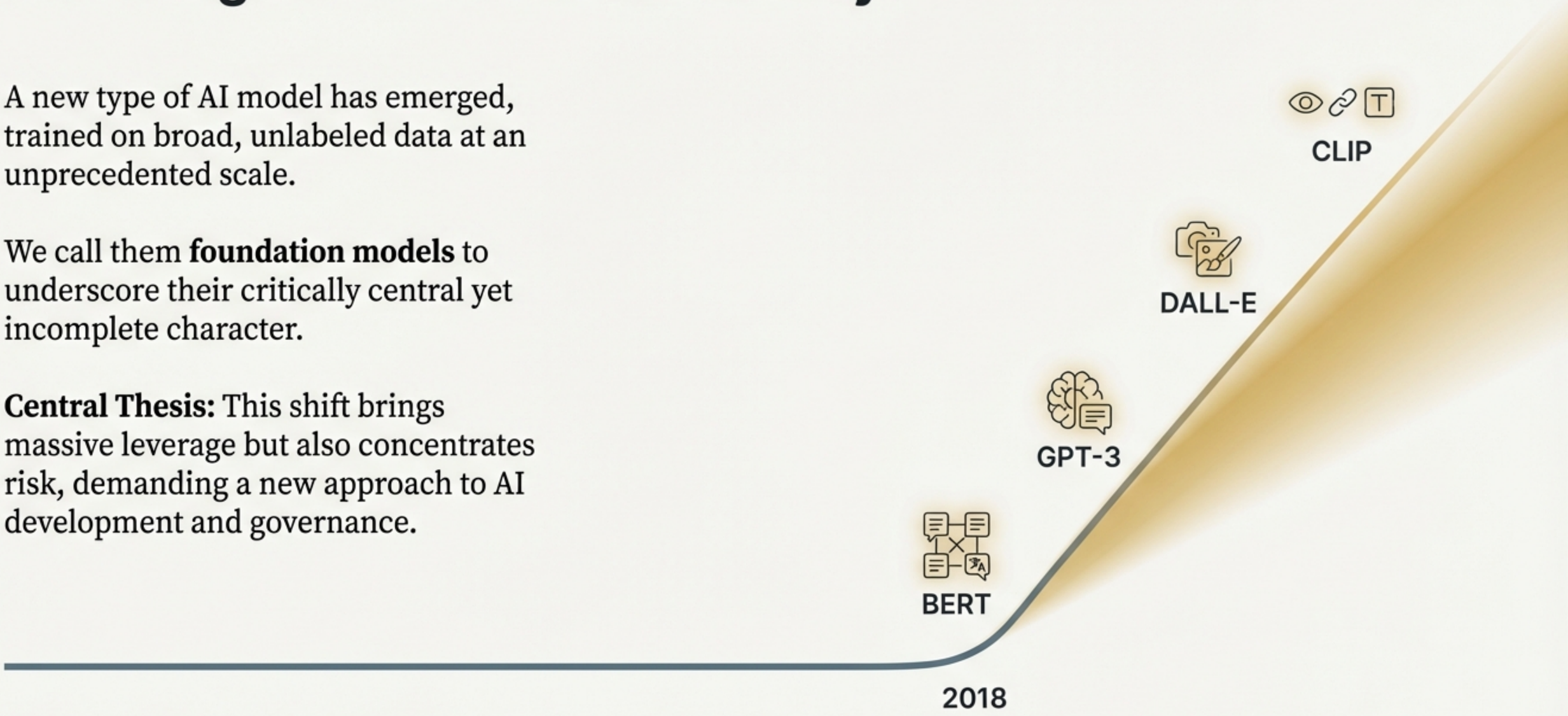


A Paradigm Shift in AI is Underway

A new type of AI model has emerged, trained on broad, unlabeled data at an unprecedented scale.

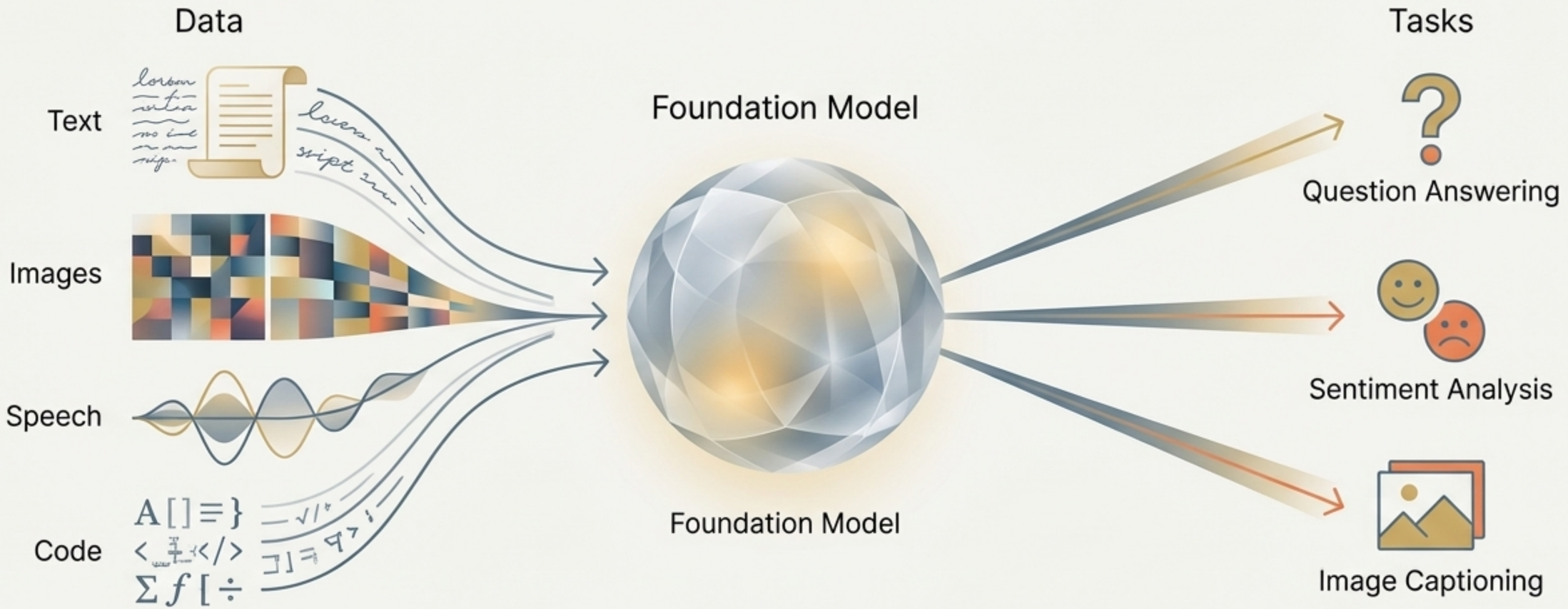
We call them **foundation models** to underscore their critically central yet incomplete character.

Central Thesis: This shift brings massive leverage but also concentrates risk, demanding a new approach to AI development and governance.



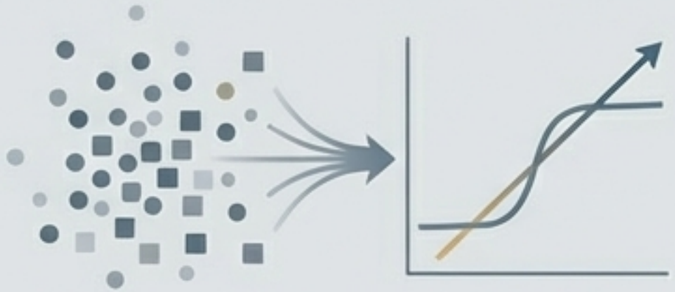
One Model, Trained Broadly, Adapted Widely

Definition: A model trained on broad data (generally using self-supervision at scale)...
...that can be adapted (e.g., fine-tuned or prompted)...
...to a wide range of downstream tasks.



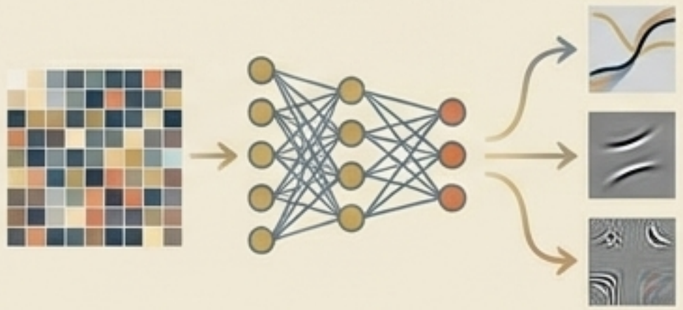
The Story of AI:
A March Towards Emergence and Homogenization

Machine Learning



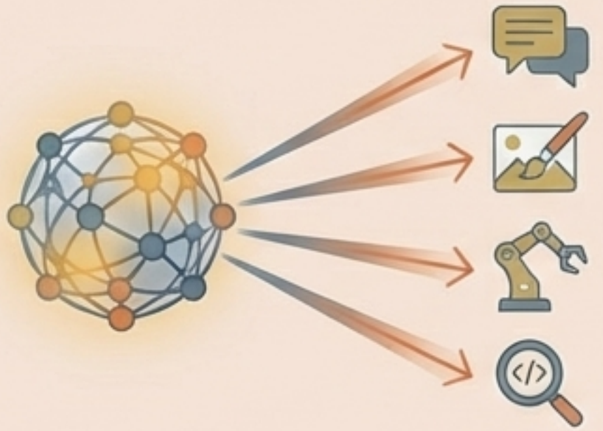
Algorithms (like logistic regression) became homogenized. The "*how*" of a task emerged from data.

Deep Learning



Architectures (like CNNs) became homogenized. The *features* used for prediction emerged from raw data (e.g., pixels).

Foundation Models



The model itself (like GPT-3) is homogenized. Advanced *functionalities* (like in-context learning) emerge from scale.

Emergence of... how → features → functionalities

Homogenization of... learning algorithms → architectures → models

The Promise: Emergence Creates Unforeseen Power

- Emergence: Behaviors are implicitly induced rather than explicitly constructed.
- Scale is the key driver.
- GPT-3 (175 billion parameters) parameters) displays in-context learning, a capability not present in smaller models and not explicitly trained for.
- This is the source of scientific excitement—models are developing capabilities that surprise even their creators.

The Peril: Homogenization Creates Systemic Risk

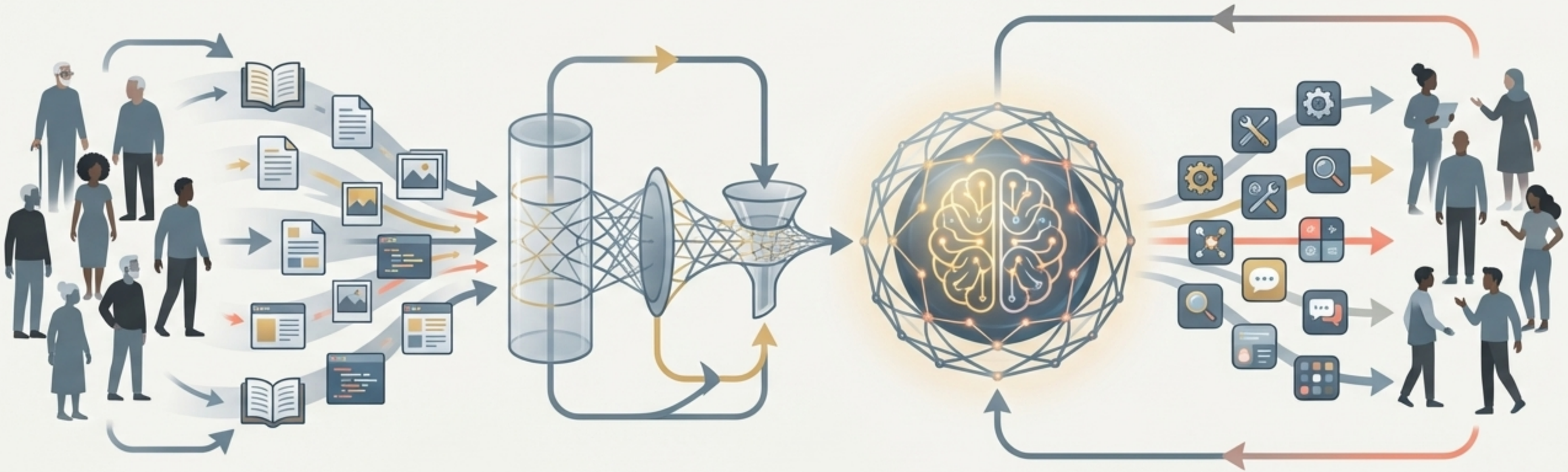
Homogenization:
Consolidation of a few models as the starting point for thousands of applications.

This creates powerful leverage: any improvement in the foundation model benefits all downstream tasks.

But it's also a liability: it creates a **single point of failure**. Any flaws, biases, or security vulnerabilities... are inherited by every application built upon it.

The Broader Ecosystem Matters

A foundation model is not an isolated artifact. Its impact is shaped by a full socio-technical ecosystem.



Data Creation

Fundamentally human, reflecting societal structures and biases.

Data Curation

Selection and filtering choices are critical but often opaque.

Training

The celebrated, but only one, part of the process.

Adaptation & Deployment

Where harms are often mitigated—or amplified.

Vast Opportunities Across Society

Healthcare & Biomedicine

Accelerating drug discovery and personalizing medicine by integrating multimodal data (images, text, genomics).

Law

Improving access to justice by automating document review, legal research, and argument generation.

Education

Enabling personalized learning at scale, creating interactive tutors, and generating educational content.

Significant Risks to Society

Inequity & Fairness

Models can amplify historical biases present in web-scale data, leading to unfair outcomes for marginalized groups.

Misuse

High-quality generated content (text, images, code) can be weaponized for disinformation, harassment, and cyberattacks.

Environmental Impact

Training a single large foundation model requires enormous computational resources, contributing to a significant carbon footprint.

Concentration of Power

Immense training costs concentrate the power to build and control foundation models in a few, very large organizations.

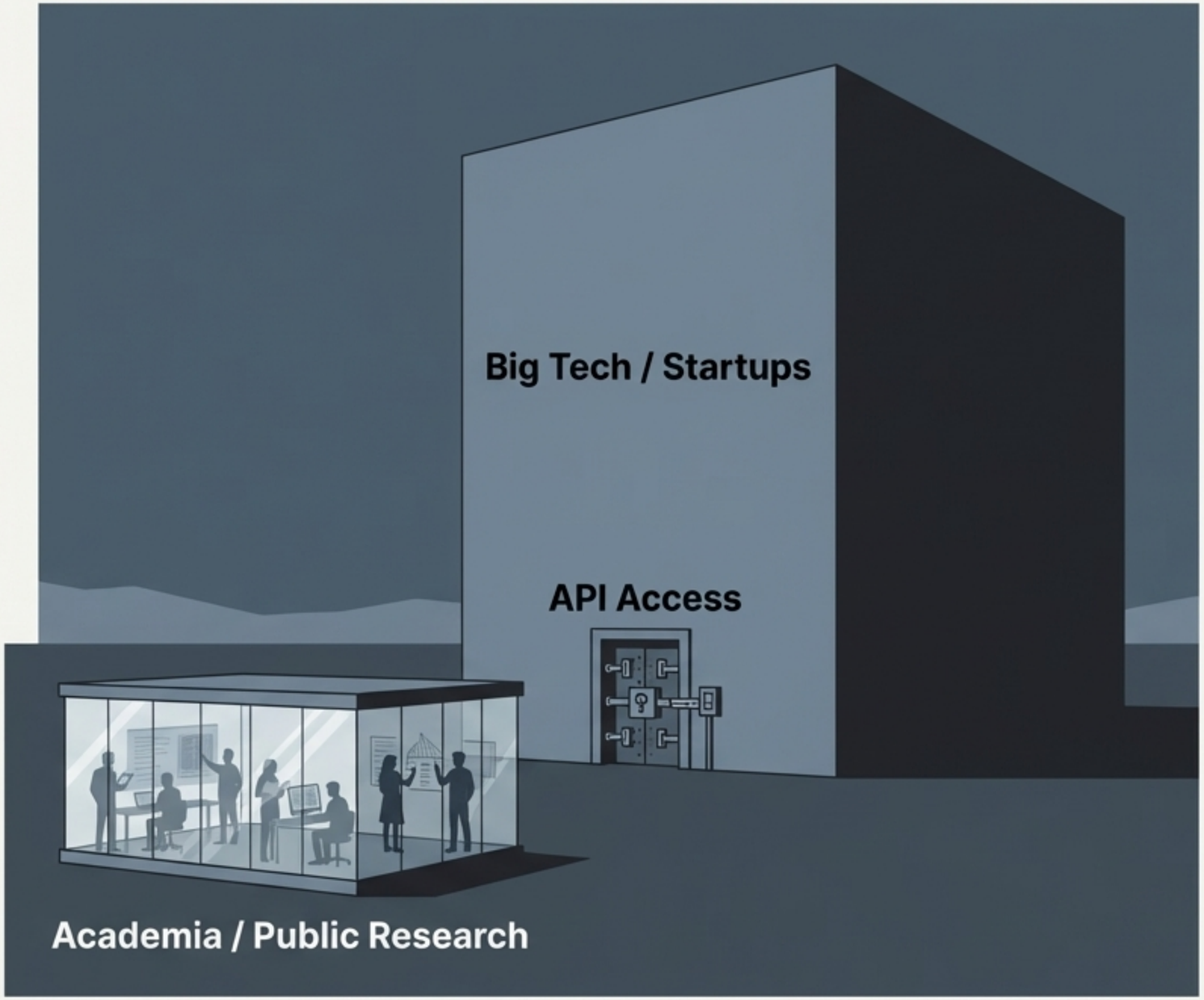
An Accessibility Crisis is Stifling Research

The computational cost to train foundation models is beyond the reach of most academic and public institutions.

Some state-of-the-art models (like GPT-3) are not released, only accessible via restrictive APIs.

This rolls back a decade of progress in open and reproducible AI research.

It prevents independent auditing and concentrates the direction of the field within industry.

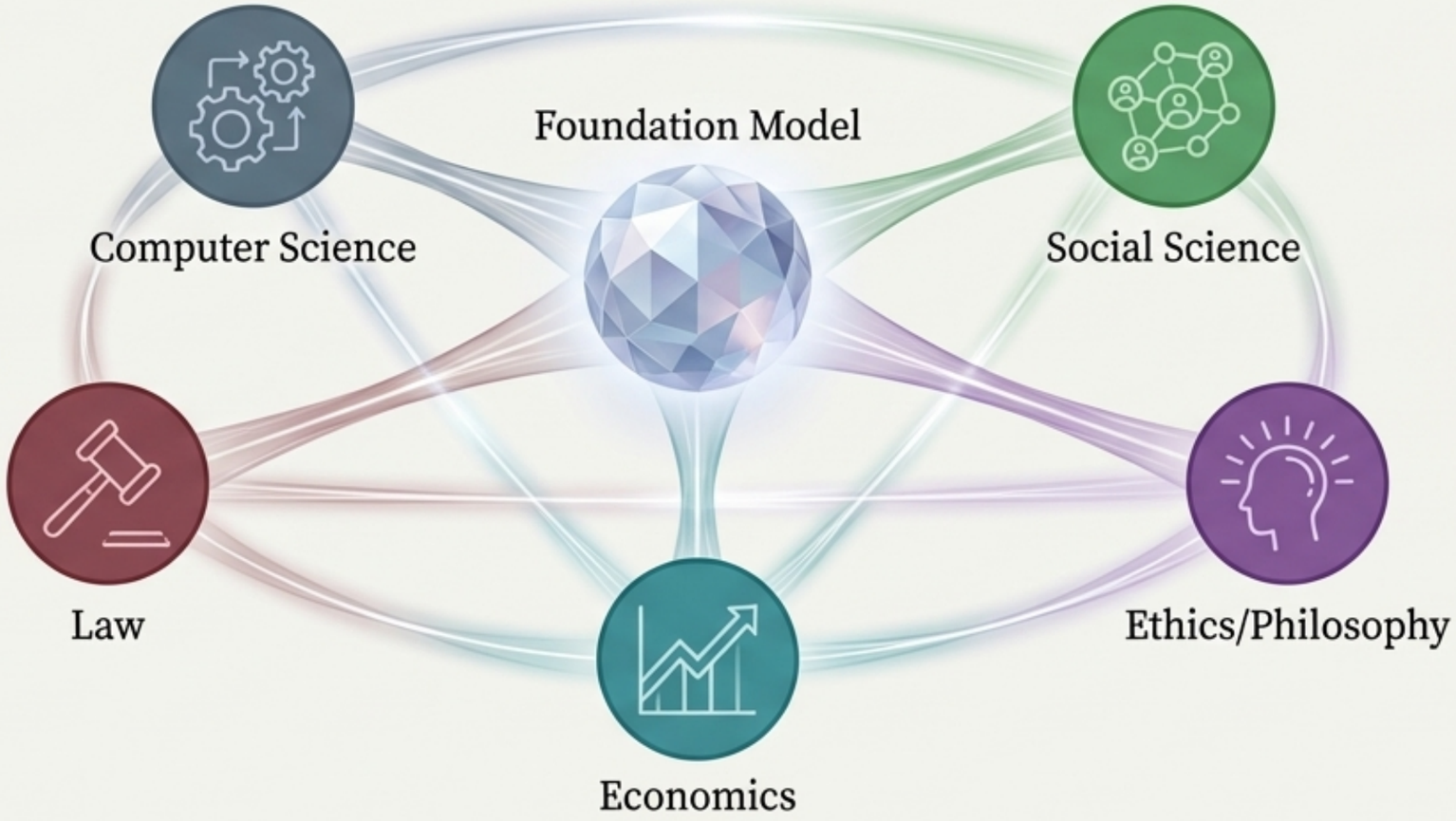


A Call for a New Science of Foundation Models

Despite their widespread deployment, we lack a clear understanding of **how foundation models work, when they fail,** and what they are truly **capable of.**

Addressing these questions requires more than just **computer science.**

We need **deep, interdisciplinary collaboration** to tackle their fundamentally socio-technical nature.



The Path Forward: A Call for Collective Action

Invest in Public Infrastructure

Create a “National Research Cloud” to democratize access to large-scale computation, similar to investments in Big Science projects.

Establish Professional Norms

Develop community standards for responsible development, release strategies, and documentation (e.g., data sheets, model cards).

Foster Interdisciplinary Research

Support and incentivize the deep collaboration needed to guide the development and deployment of foundation models for societal benefit.