# The Second Curve of Scaling Law

aka.ms/**GeneralAI**

Furu Wei, Partner Research Manager

Microsoft Research

Dec. 2023

# The New Age of AI (2020/GPT-3 - )

## Paradigm Shift 2022/ChatGPT

"What I cannot create, I do not understand."

Richard Feynman Generative models (openai.com), June 16, 2016

**Generative Pre-training**

4 **In-Context Learning (Zero/Few-Shot Learning)**

Natural language prompt (instructions) →

Context/Input →

Demonstrations (optional) →

System prompt (optional) →

5 **Reasoning Engine**

**Foundation Model / (M)LLMs**

*(billions/trillions Parameters)*

Example(s): GPT-4/4V, …

→ Output: Generate anything: text, image, audio, video, …

3 **Generative AI**

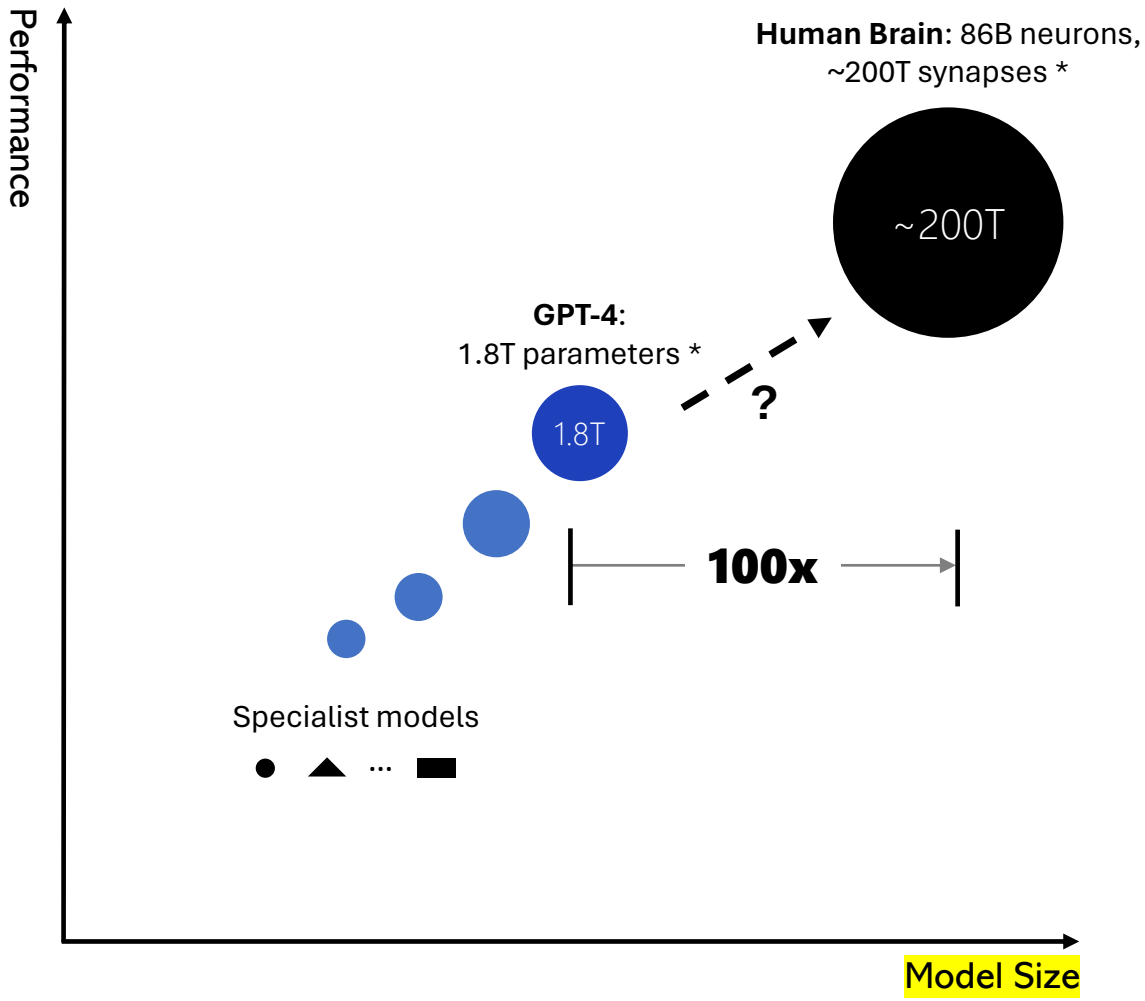
2 **Language Interface**

**One (Very Large General) Model**

Across tasks, languages, and modalities

1 **General AI**

# The Power of Scaling (Law)

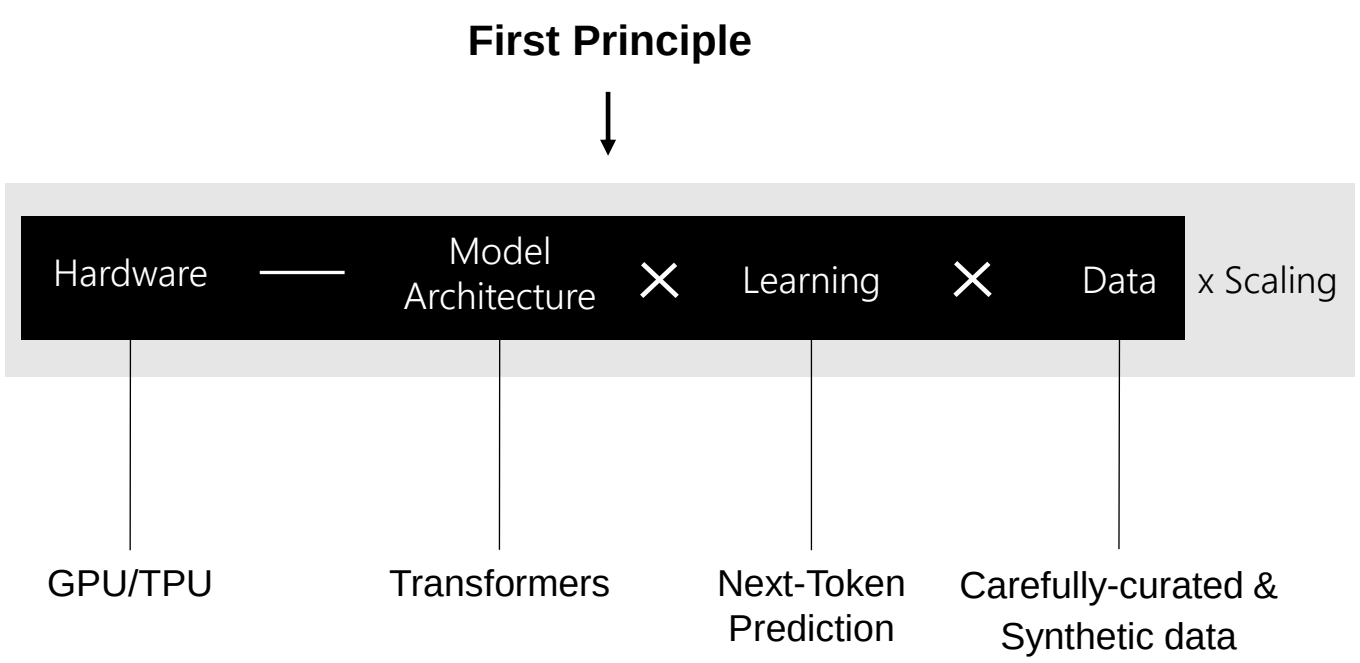


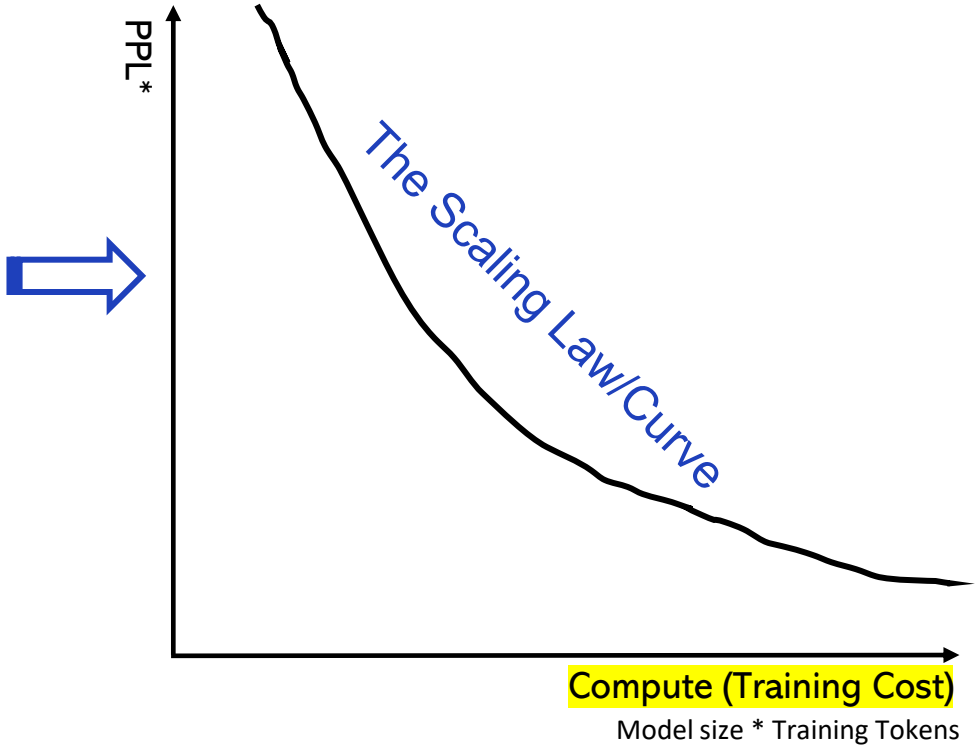

**Is (the first) scaling (curve) all we need?**

**YES and NO**

* GPT-4 architecture, datasets, costs and more leaked
* List of animals by number of neurons - Wikipedia

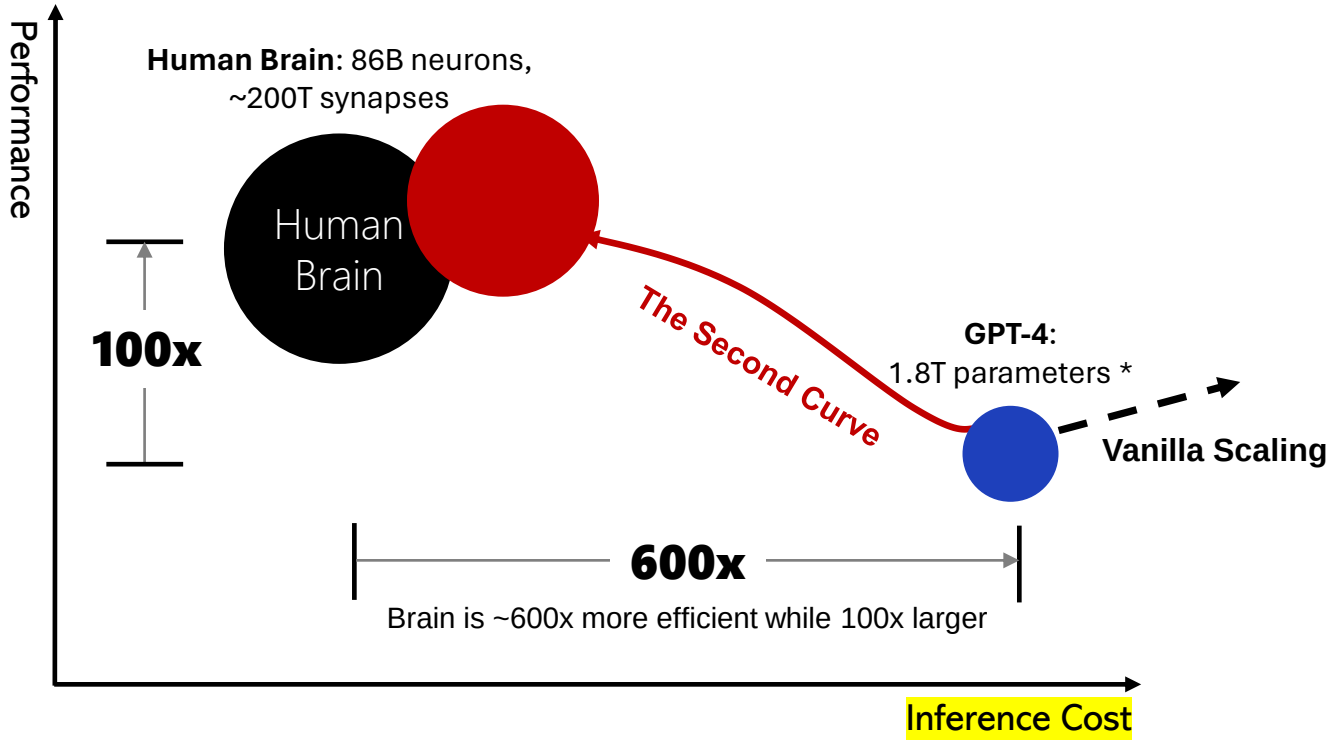
# First Principle of Scaling Law (Curve)

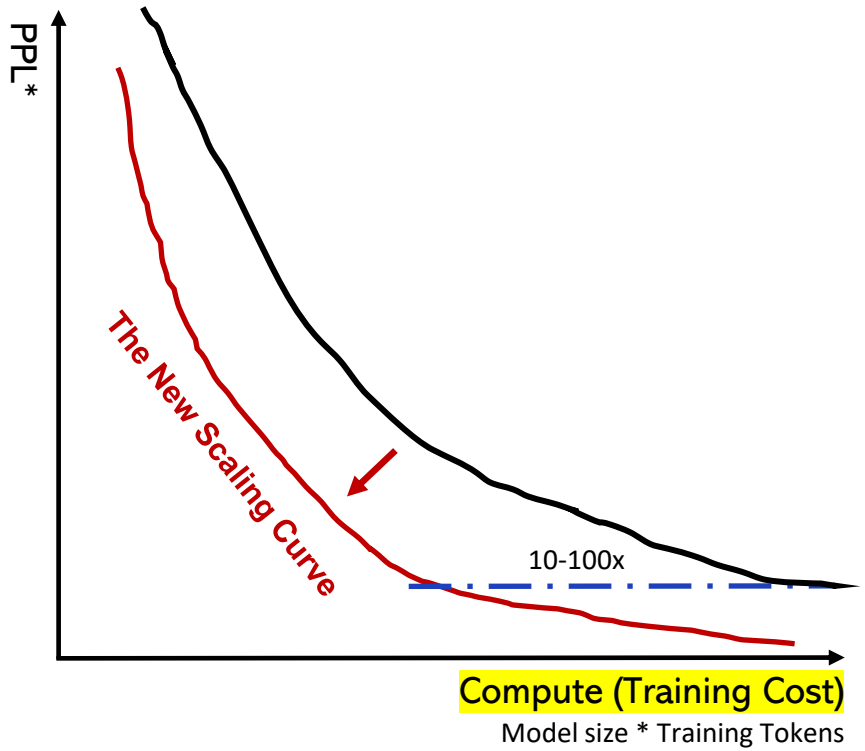



* Lower PPL = higher compression ratio / performance / intelligence

# The Second Curve of Scaling Law

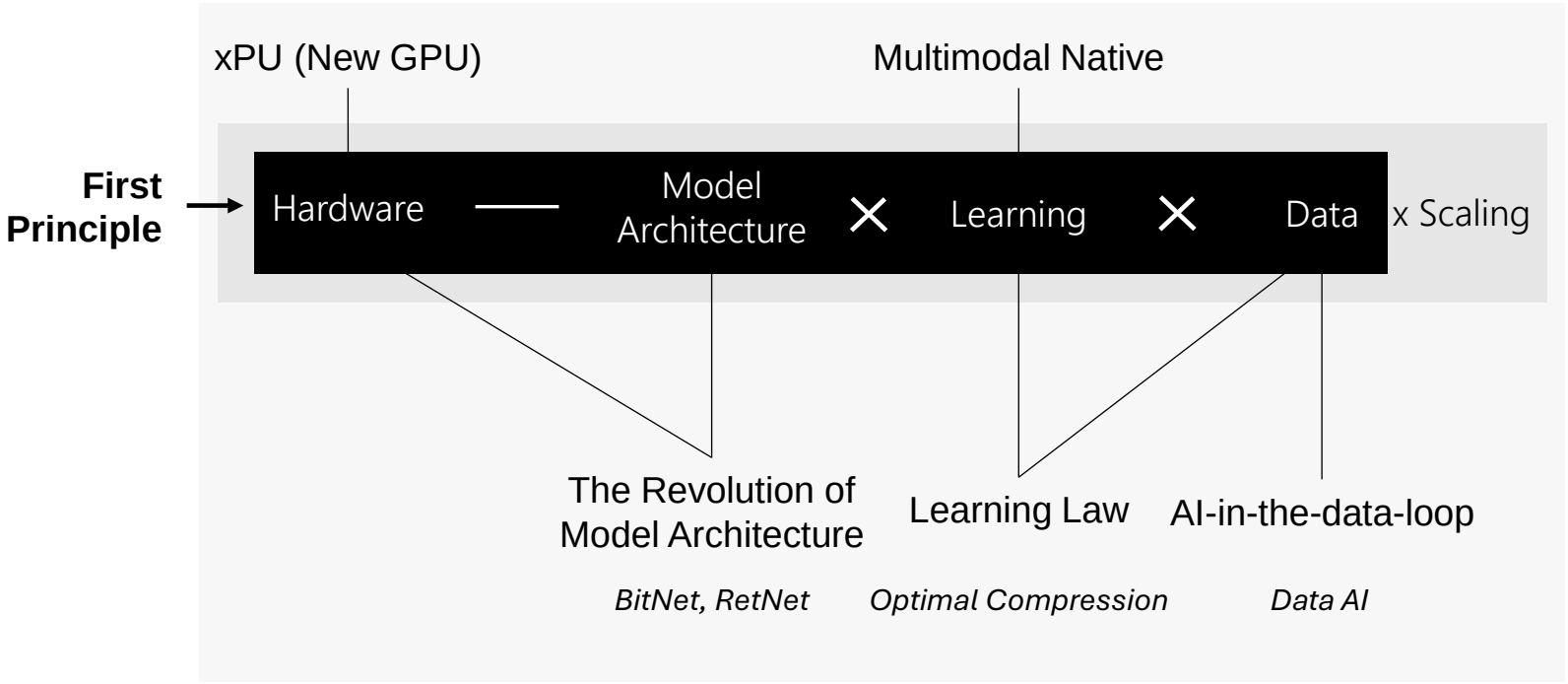


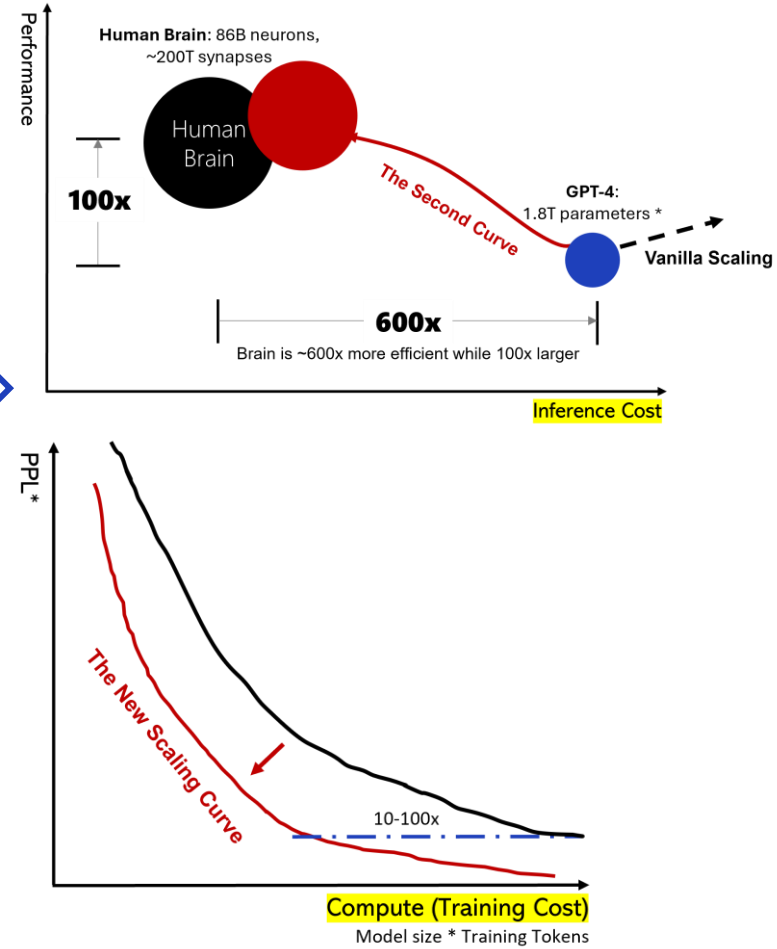




* [GPT-4 architecture, datasets, costs and more leaked]
* [List of animals by number of neurons - Wikipedia]
* [Does Thinking Really Hard Burn More Calories? - Scientific American]

* Lower PPL = higher compression ratio / performance / intelligence

# The Second Curve of Scaling Law



aka.ms/GeneralAI

Contact: Furu Wei (fuwei AT microsoft.com)