The Second Curve of Scaling Law

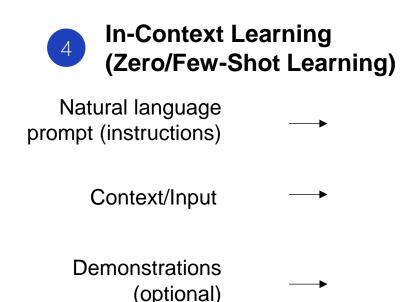
aka.ms/**GeneralAl**

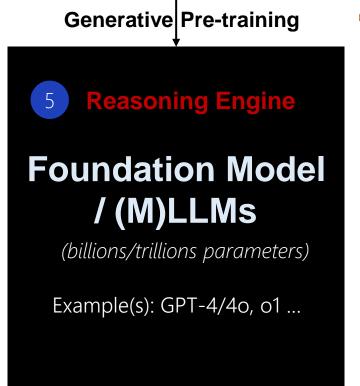
Furu Wei, Partner Research Manager
Microsoft Research
Dec. 2023

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 Generate anything: text, Output image, audio, video, ... **Generative Al**

2 Language (Multimodal) Interface

System prompt

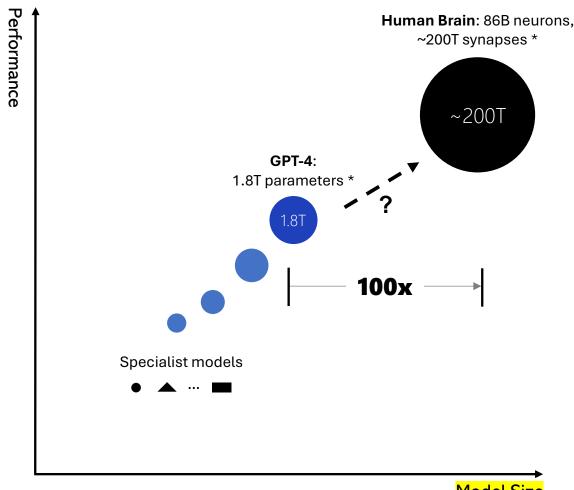
(optional)

One (Very Large General) Model

Across tasks, languages, and modalities

General Al

The Power of Scaling (Law)



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

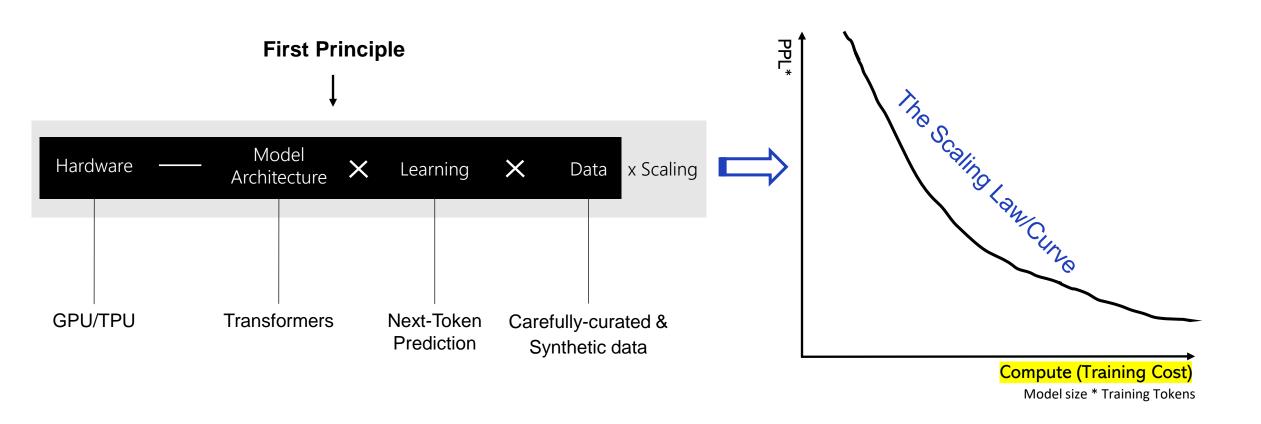
YES and **NO**

Model Size

^{*} GPT-4 architecture, datasets, costs and more leaked

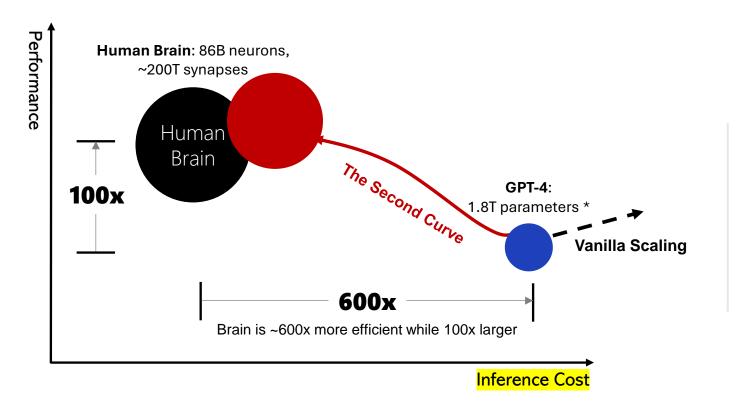
^{*} List of animals by number of neurons - Wikipedia

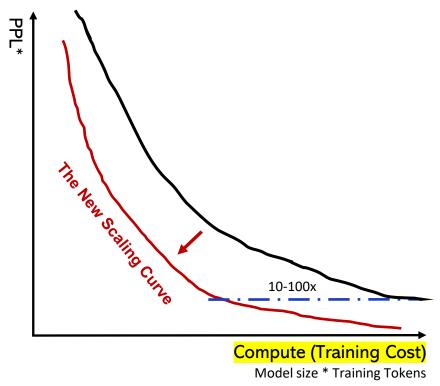
First Principle of Scaling Law (Curve)



^{*} Lower PPL = higher compression ratio / performance / intelligence

The Second Curve of Scaling Law





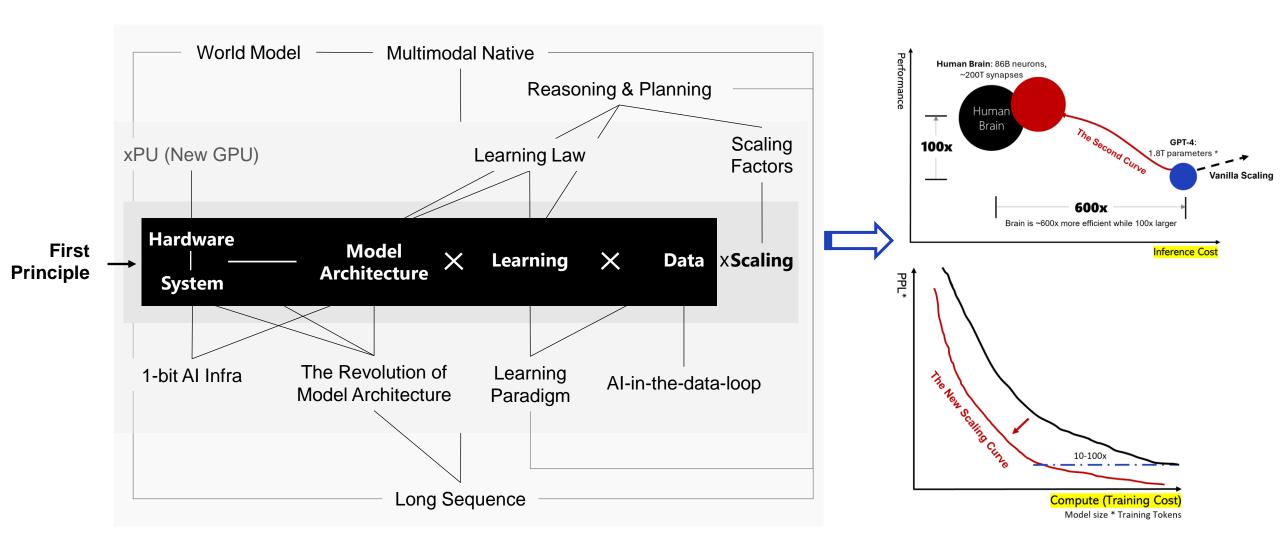
^{*} GPT-4 architecture, datasets, costs and more leaked

^{*} List of animals by number of neurons - Wikipedia

^{* &}lt;u>Does Thinking Really Hard Burn More Calories?</u> - <u>Scientific American</u>

^{*} Lower PPL = higher compression ratio / performance / intelligence

The Second Curve of Scaling Law



aka.ms/**GeneralAI**

Contact: Furu Wei (fuwei AT microsoft.com)