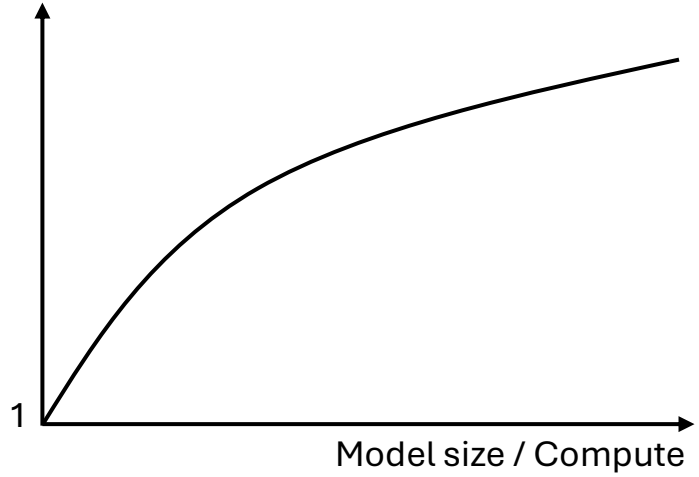
Scaling Factors

Scale is the basic principle under LLMs and future AI. What we need are **scaling-friendly factors**.

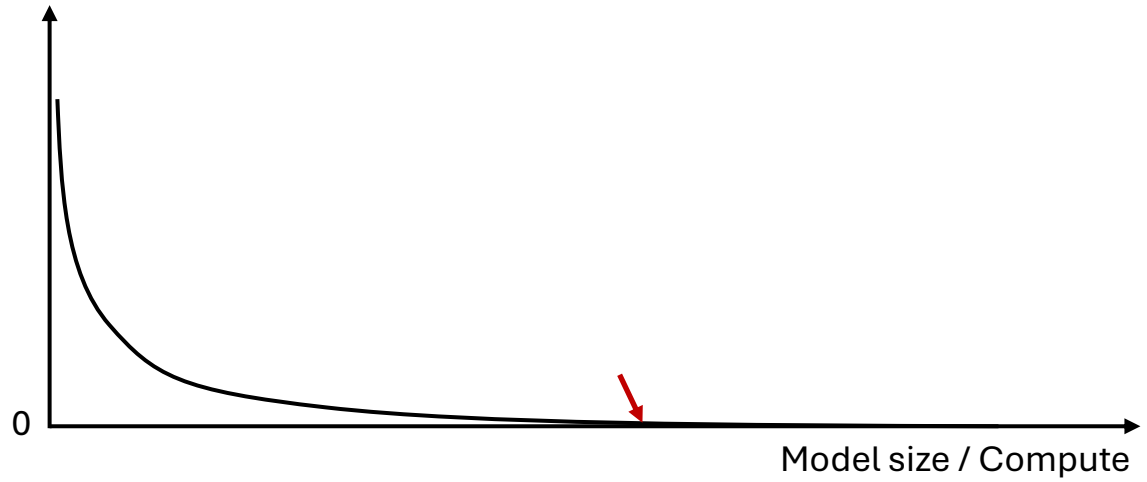
The **advantage** of a technique or research can be **amplified** (not diminished) by scaling up model (and/or data) size

The **disadvantage** of a technique or research (w/ other significant benefits) can be **diminished** by scaling up model (and/or data) size

$$\text{Gain} = \frac{New}{Baseline}$$

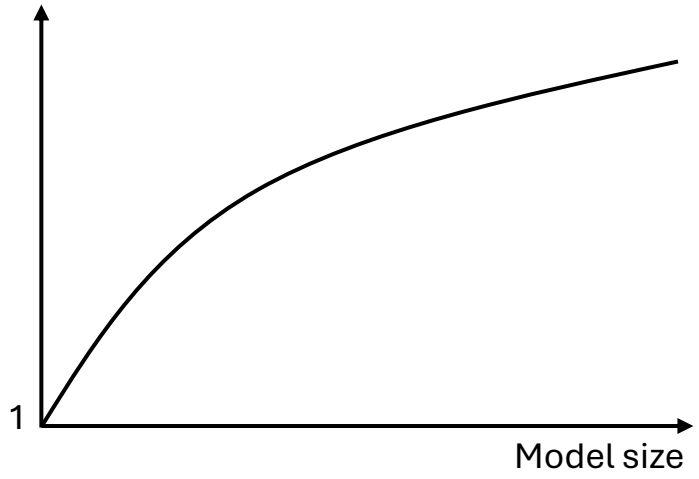


Gap = Baseline – New



An Example: 1-bit LLM / BitNet

$$\text{Efficiency gain} = \frac{\text{FP16}}{\text{1bit}}$$



Accuracy gap = FP16 – 1bit

aka.ms/GeneralAI

Contact: Furu Wei (fuwei)