AMITY UNIVERSITY

Large Language Models (LLM)

- Module 1
- Lecture 4

  Scaling Laws and Model Efficiency

What Are Scaling Laws?

- Scaling Laws describe how model performance improves as we increase:
  - Model size (parameters)
  - Dataset size (tokens)
  - Compute budget (training steps)
- Introduced in papers like Kaplan et al. (2020) — "Scaling Laws for Neural Language Models.“
- They provide a predictable roadmap for training larger and better LLMs.
- Why Important?
  - Helps design future LLMs efficiently without random guessing.

Observations from Scaling Laws

- Performance improves smoothly and predictably as:
  - ➢ Parameters ↑
  - ➢ Data ↑
  - ➢ Compute ↑
- Power-law relationships: Error = A × (Resource)^-α
- No sudden saturations — gradual improvement.
- Error/loss decreases following a power law as size and data increase.
- Earlier trend: "Bigger models = Better models.“
- Corporations scaled models massively (GPT-2 → GPT-3 → Gopher).
- Problem: Oversized models trained on insufficient data were undertrained.

Key Question

"Given a fixed compute budget, how should we balance model size and dataset size to get the best model?

"You can't simply keep growing parameters without growing data proportionally!

Model efficiency

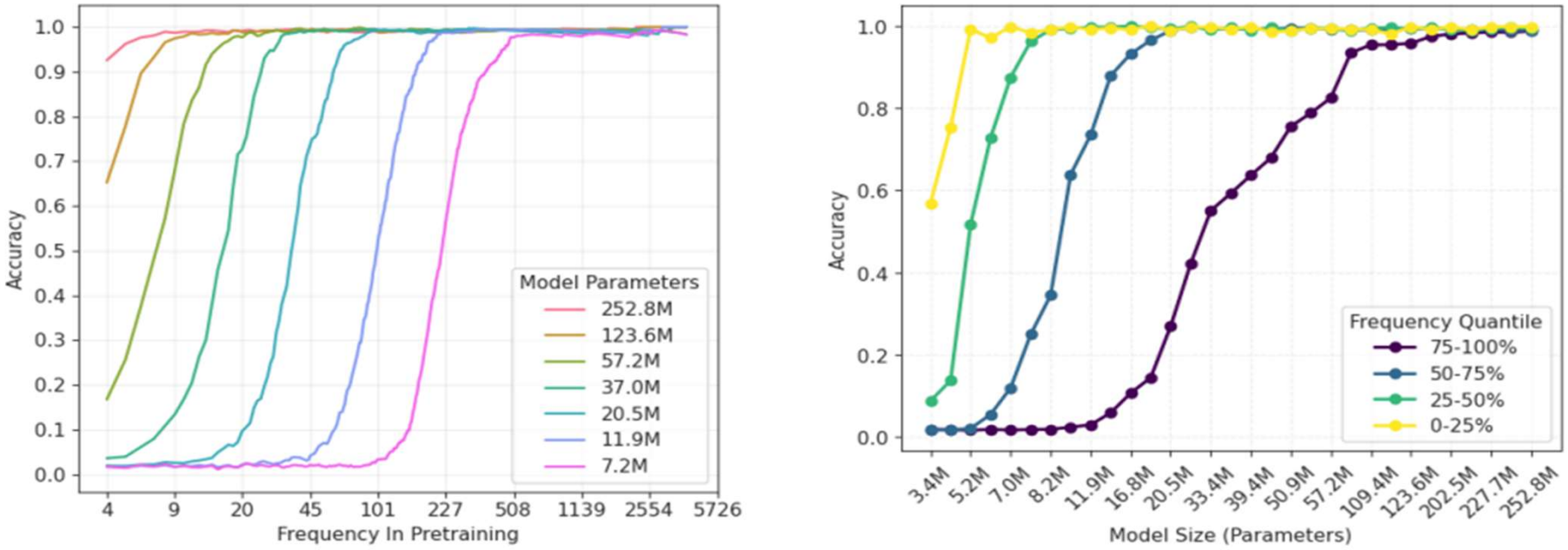


Figure 2: (a) Accuracy of sufficiently trained models with different sizes across varying input frequencies. When the frequency falls below a model-specific threshold, small models inevitably hallucinate and fail to learn the corresponding facts. (b) Accuracy of different frequency classes (split into four quantiles) under varying model sizes. As model size increases, the model progressively learns the more frequent data first, while infrequent data becomes learnable only at larger scales.

Kaplan Scaling Laws (2020)

- Finding: For increased compute, scale model size more than data.
- GPT-3 example:
  - 175B parameters
  - Only 300B tokens (~1.7 tokens per parameter)
- Problem: Undertrained at that size.

Chinchilla Scaling Laws (DeepMind, 2022)

- Correction: Scale model size and dataset size equally with compute.
- Findings:
  - Data more important than previously thought.
  - GPT-3 was severely undertrained.
  - For optimality: ~20 tokens per parameter.

The Chinchilla Trap

- Problem: Chinchilla-optimal models are efficient for training but expensive at inference.
- Example: LLaMA-1, LLaMA-2, LLaMA-3 models.
  - Extremely high token-to-parameter ratios (e.g., 1875 tokens/parameter in LLaMA-3).
- New Trend: Train smaller models longer for cheaper inference!

Scaling Test-Time Compute (2024–2025)

- New paradigm:
  - ➢ Instead of larger models only, use more compute at inference.
- Chain-of-Thought Reasoning: Give models time to "think" (multi-step answers).
- RL at Test Time: Reinforcement learning during answer generation (OpenAI's o1, o3, Gemini Flash).

New Types of Scaling Laws

Category	Strategy	Example Models
Pre-training scaling	Train on more tokens	Chinchilla, LLaMA 2
Post-training scaling	Use synthetic data	GPT-4o, Gemini Ultra
Test-time scaling	More inference steps	OpenAI o1, o3, Gemini Flash

Modern Efficiency Techniques

- Mixture of Experts (MoE): Activate small parts of the network (Mixtral, DeepSeek).
- Quantization: Compress models to 8-bit/4-bit.
- Distillation: Smaller "student" models inherit from larger "teacher" models.
- Long Context Training: Models handling 128k–1M tokens (Claude 3, Gemini 2).

Real World Example: LLaMA-3 Scaling

- 8B model trained on 15T tokens (1,875 tokens/parameter!).
- Outperforms older larger models at same inference cost.
- Training longer = smaller, faster, cheaper models.

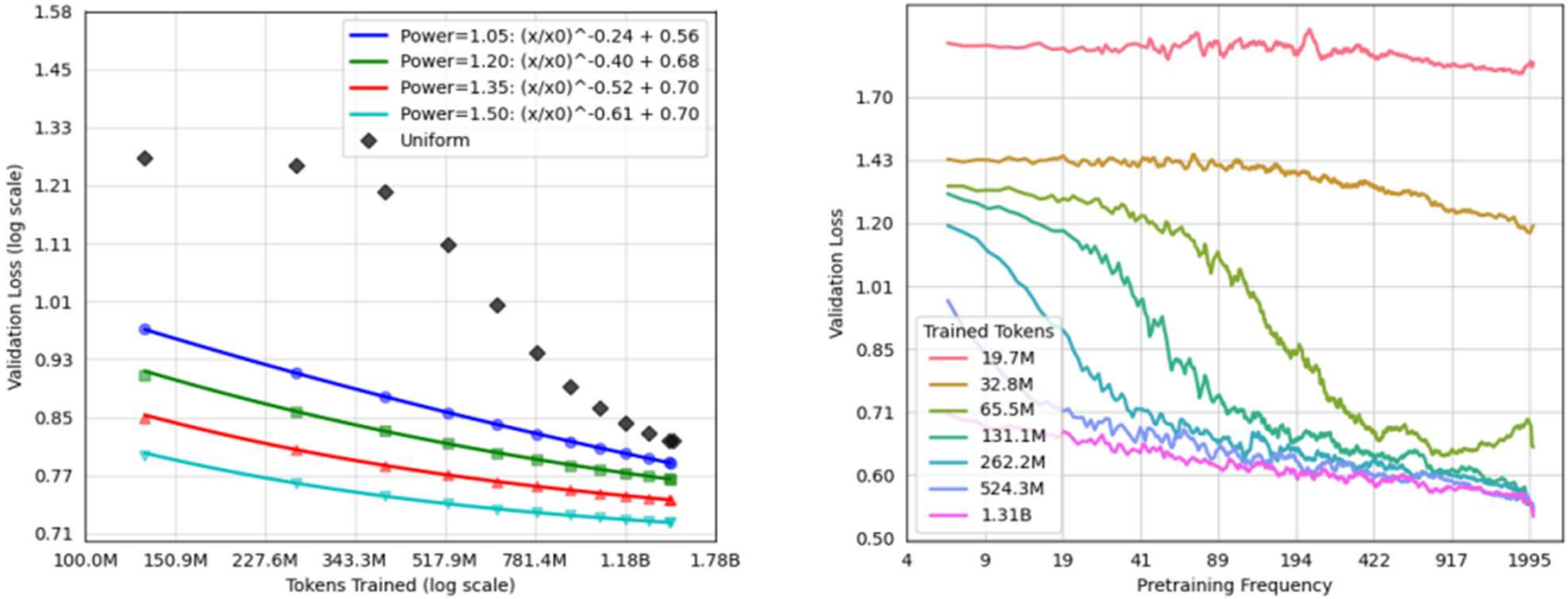


Figure 4: (a) Validation loss as a function of training data size. Models trained on data sampled from pretrained knowledge under various power-law distributions (i.e., $p(i) \sim (x + b)^{\text{power}}$) show clear power-law scaling of loss with data size, while uniform sampling does not. A more skewed data distribution leads to faster loss decay. (b) Loss decomposition by data frequency class shows a frequency-dependent learning pattern: high-frequency data is learned earlier, while lower-frequency data is acquired later during training.

Challenges in Modern Scaling

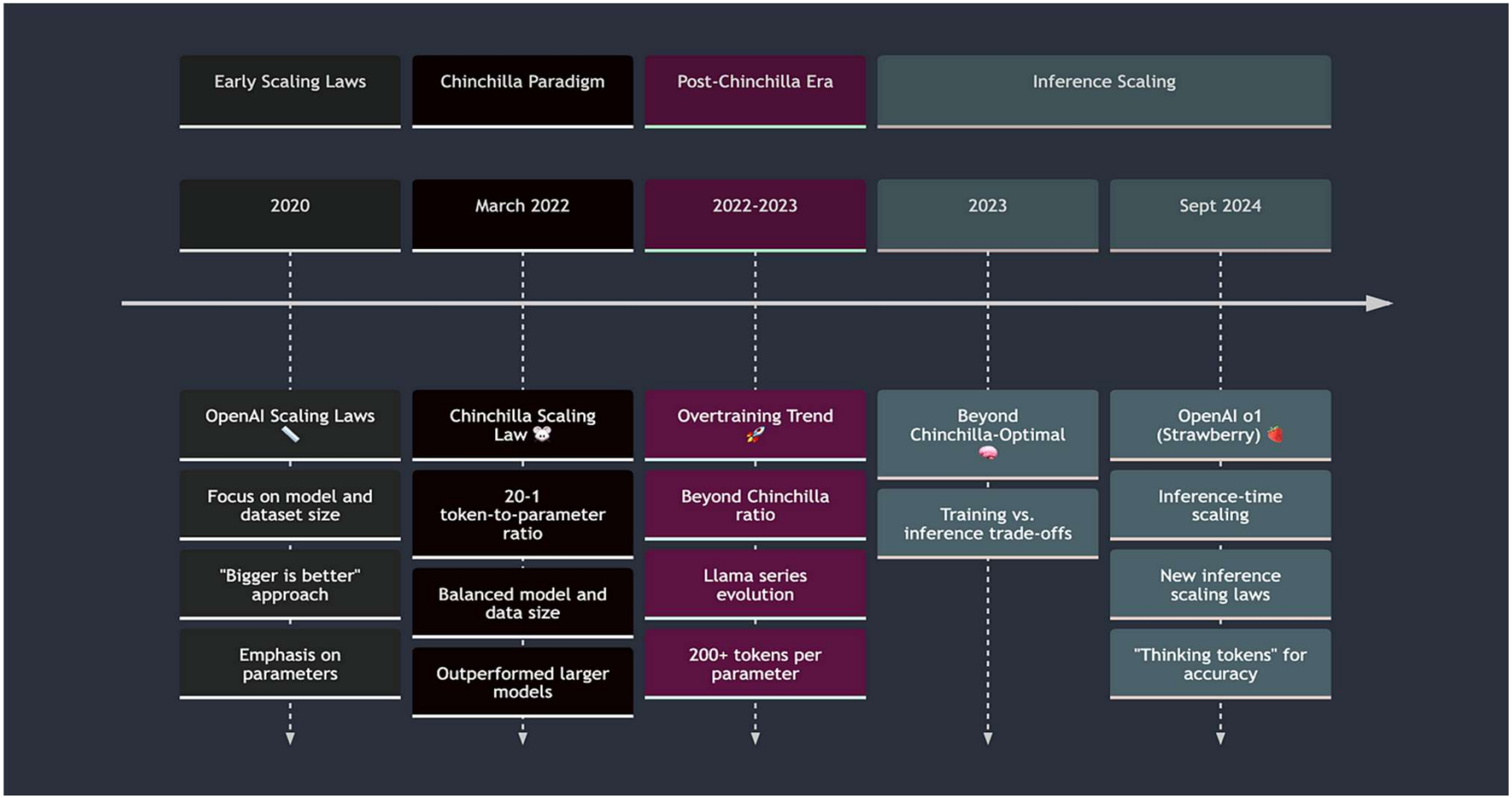
- Training Compute Explosion:
  - Gemini Ultra trained with 5×10^{25} FLOPs.
- Inference Costs:
  - Even efficient models are expensive at deployment scale.
- Data Bottleneck:
  - Limited access to new, clean, high-quality datasets.

Future of LLM Scaling Beyond 2025

- Pretraining: Limited due to data saturation.
- Post-training (Synthetic Data): Major growth area.
- Inference-time scaling: Models "think more" instead of growing bigger.
- Smarter, not bigger: New focus on reasoning and cognitive scaling.

Take-away points

- Scaling Laws have evolved:
  - From size + compute → to data + inference + reasoning scaling.
- The next generation of LLMs (2025–2030) will prioritize:
  - Efficient learning
  - Cheaper inference
  - Smarter, high-reasoning AI


Early Scaling Laws
Chinchilla Paradigm
Post-Chinchilla Era
Inference Scaling
2020
March 2022
2022-2023
2023
Sept 2024
OpenAI Scaling Laws
Chinchilla Scaling Law
Overtraining Trend
Beyond Chinchilla-Optimal
OpenAI o1 (Strawberry)
Focus on model and dataset size
20-1 token-to-parameter ratio
Beyond Chinchilla ratio
Training vs. inference trade-offs
Inference-time scaling
"Bigger is better" approach
Balanced model and data size
Llama series evolution
New inference scaling laws
Emphasis on parameters
Outperformed larger models
200+ tokens per parameter
"Thinking tokens" for accuracy