Parameters: From Millions to Trillions in GPT, Llama and Deepseek

Parameters in AI Models

- Parameters are core components in AI models, often referenced by their total count
- Early machine learning models had 20–200 parameters
- Modern models have exploded in size: GPT-1 (117M), GPT-2 (1.5B), GPT-3 (175B), GPT-4 (1.76T)
- Some current models' parameter counts are undisclosed, but likely tens of trillions
- Efficiency improvements mean smaller models (e.g., Gemma at 270M) can outperform older, larger ones
- Generally, more parameters = more intelligence, more training data absorbed

Model Variations and Scaling

- Models often come in different sizes (e.g., GPT-5 Nano, Mini, Full) with increasing parameter counts
- Claude has Haiku (small), Sonnet (medium), Opus (large)
- Larger models cost more to run due to higher compute needs
- Training time scaling: bigger models, more parameters, more training data, higher cost
- Chinchilla scaling laws: model size should match training data volume for best results

Inference Time Scaling

- Inference: using the model after training
- Techniques can improve model performance at inference time (not just by making models bigger)
- Examples: prompting the model to reason step-by-step, providing richer input sequences
- Retrieval-Augmented Generation (RAG) is a key inference time scaling method
- Both training time and inference time scaling are now important for model performance

Trends and Examples

- Shift in last year: more focus on inference time scaling, not just bigger models
- Open source models: Llama 3.2 (3B), Llama 3.3 (3.3B), Llama 4 (245B multimodal), DeepSeek (671B)
- Most large models now use "mixture of experts" architectures—multiple smaller models inside one system
- Parameter counts often shown on logarithmic scales due to huge differences in size

Additional Resources

- Short videos explaining parameters are available in the provided resources
- Encouraged to review these if unfamiliar with the concept of parameters