Understanding Transformers: The Architecture Behind GPT and LLMs

Course Progress & Capabilities

- Day four of the course, lots of new content planned
- Already able to write code to call OpenAI in the cloud and Llama locally
- Can compare strengths and limitations of advanced AI models
- Will cover topics like agency in AI, context engineering, agent loops, and hot trends
- Basics like tokens, context windows, parameters, API costs to be reviewed
- Aimed at both beginners and experienced folks—something for everyone

Competition Results & Model Rankings

- Recap of a recent AI competition called "outsmart"
- Winner: Charlie (played by Grok4), known for being devious
- Close scores: Charlie first, OSS second, GPT5 third, Drew last, Claude Sonnet at the bottom
- Drew led for a while, but others ganged up, changing the outcome
- Over 2,000 games played overall; Grok strong but with only three games
- Claude 3.5 Sonnet performed best in terms of most games played
- Rankings and code available on the main site; blog post explains tactics and game variants
- Users encouraged to try the game, check rankings, and experiment locally

Transformer Models & AI History

- GPT stands for Generative Pre-trained Transformer
  - G: Generative—predicts next tokens in a sequence
  - P: Pre-trained—trained on large internet datasets
  - T: Transformer—refers to the model architecture
- Focus on practical learning—understanding through code and hands-on examples
- Will gradually explore transformer internals over the next eight weeks

The Story of the Transformer

- Transformers introduced in 2017 by Google’s "Attention Is All You Need" paper
- Paper described a new neural network architecture, focusing on self-attention layers
- Self-attention lets models focus on important parts of input sequences
- Enabled larger, more efficient models and faster training on bigger datasets

Neural Networks Background

- Traditional data science: statistical models predict outcomes from parameters
- Neural networks: inspired by the brain, use many interconnected artificial neurons
- Deep learning: stacking layers for deeper pattern recognition
- History of ups and downs, breakthroughs allowed bigger, deeper networks

Evolution of GPT & LLMs

- OpenAI released GPT-1 (basic), then GPT-2 (improved), then GPT-3 (major leap)
- ChatGPT (2022) used GPT-3.5 and RLHF (reinforcement learning from human feedback) for chat mode
- GPT-4 (2023) brought multimodal capabilities, now GPT-5 and possibly GPT-6
- Rapid, surprising progress in the field, even for industry insiders

Transformer Architecture: Importance & Alternatives

- Transformer is not fundamentally required for token prediction—it's an optimization
- Main value: efficiency, scalability, lower API costs
- Without transformers, progress would have been slower and more expensive
- Alternatives exist (state space, hybrid architectures), but none have surpassed transformers yet
- Transformer remains the main architecture for large language models, but the field is evolving

Next Steps

- More in-depth exploration of transformer internals in upcoming sessions
- Encouraged to experiment with model games, rankings, and code on the course website