From LSTMs to Transformers: Attention, emergent Intelligence and Agentic AI

Key Discussion Points

- LSTM (Long Short Term Memory) was the main model before Transformers
- LSTM seen as powerful, deeper sequence understanding, but hard to parallelize
- Training LSTMs slow and complex due to step-by-step learning
- Transformers simplified things, focus on parallelization and scalability
- "Attention is All You Need" paper argued attention alone is enough, enabling huge performance gains
- Transformers' success surprised the world in 2023, widespread public interest
- Initial excitement followed by backlash, concerns about "stochastic parrots" (predicting likely words, not truth)
- Surprising outcome: large models often output not just plausible, but accurate and intelligent responses
- Emergent intelligence: with enough scale, models simulate intelligent responses, still not fully understood

Industry Trends and Innovations

- Prompt engineering briefly became a high-paying job, now a common skill
- Copilots (e.g., Microsoft Copilot, GitHub Copilot) enable human-AI collaboration, automate and enhance work
- Shift from prompt engineering to context engineering—focus on supplying LLMs with rich, relevant input
- Context engineering includes providing business-specific info, tools, and structured prompts for better outputs
- Example: including ticket prices in prompts to ensure accurate customer info

Agentic AI

- Agentic AI is a leading topic, with dedicated courses and growing interest
- Two main definitions:
    - LLM controls workflow, can call other LLMs or tools
    - LLM operates in a loop, calling itself, using tools, planning and executing tasks (e.g., Claude code)
- Autonomy in agentic AI: LLMs generate instructions for next actions, appear to "choose" their own steps
- Real-world examples: Claude code, GPT agents booking reservations
- Promise of agentic AI: more autonomous, task-oriented AI systems, hands-on experience planned for course participants