AI fundamentals: Understanding key terms and ideas

Key terms

- Artificial Intelligence (AI): Creating smart machines that perform tasks like humans.
- Large Language Model (LLM): A type of AI trained on vast amounts of text to understand and generate human-like language. Claude, Bard, and GPT-4 are all examples of LLMs.
- Hallucination: When an AI model generates incorrect or nonsensical information, believing it to be true.
- Bias: A tendency of AI to favor certain outcomes or interpretations, often reflecting the data it was trained on.
- Ethics: Principles that govern the responsible creation and use of AI, focusing on fairness, privacy, and transparency.
- RLHF (Reinforcement Learning from Human Feedback): A training method where AI models are refined based on human feedback to improve their responses and decisions.

What is AI?

- Machine learning, deep learning, and generative AI are all subsets of AI.
- Machine learning is a method enabling computers to learn from data, improve through experience, and make predictions or decisions.
- Deep learning is a complex form of machine learning using multi-layered neural networks to analyze and interpret large amounts of data.
- Generative AI is a type of AI that creates new content, like text, images, or music, by learning from existing data.

How do LLMs work?

- Data Training: LLMs learn language patterns from analyzing vast amounts of text data.
- Pattern Recognition: They recognize and predict language structures to generate coherent text.
- Context Understanding: They're capable of grasping context and nuances in language to provide relevant responses.
- Continuous Learning: LLMs improve over time by processing new data and user interactions.

How are these models trained?

- Pre-Training on Diverse Data: Models initially learn from a vast array of internet data, including websites and books, using extensive resources and time.
- Enhanced Understanding through Tuning: Post pre-training, models undergo instruction tuning or RLHF, focusing on improving query responses and interaction quality.
- Utilizing Human-Labeled Data: Refinement processes involve human-labeled data to enhance AI responses, without introducing new facts.
- Safety Measures in Data Selection: Harmful or inappropriate content is filtered out during data preparation, ensuring safer model outputs.
- Ongoing Updates and Audits: Regular updates based on user feedback and external audits maintain the model's alignment with ethical and safety standards.