The Prediction Machine

How Large Language Models Work—and the Ethics of AI

Brendan Shea, PhD

Programming and Problem Solving • Rochester Community and Technical College

Introduction

In November 2022, OpenAI released ChatGPT, and within two months, it had 100 million users—the fastest-growing application in history. Suddenly, anyone could have a conversation with an AI that wrote essays, explained concepts, generated code, and engaged in seemingly intelligent dialogue. The age of **Large Language Models** (LLMs) had arrived.

But what *is* ChatGPT? How can a computer program write poetry, answer questions about history, and help debug code? The answer is simultaneously simpler and more profound than most people realize. At its core, an LLM is a prediction machine: given some text, it predicts what text should come next. That's it. Yet from this simple principle emerges behavior that looks remarkably like understanding.

This case study explores how LLMs work—from the basic concepts of neural networks and training to the architecture that made modern AI possible. We'll then examine the ethical questions these systems raise: questions about bias, truth, labor, power, and what it means to create machines that mimic human thought.

A Technology of Contradictions

LLMs are simultaneously:

- **Simple in principle**: Predict the next word
- **Vast in scale**: Hundreds of billions of parameters trained on trillions of words
- **Impressive in capability**: Write, reason, translate, code
- **Limited in understanding**: No true comprehension, just sophisticated pattern matching
- **Transformative in impact**: Reshaping education, work, and information itself

Understanding both the power and the limitations is essential for using these tools wisely.

1 A Brief History of Language Models

From ELIZA to GPT-4

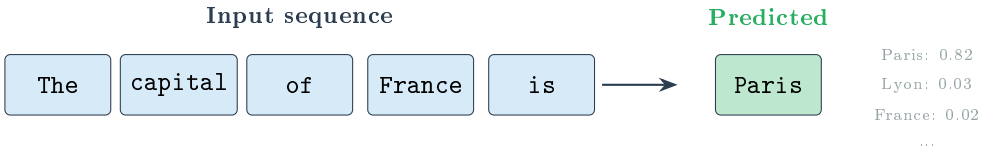
1966: Joseph Weizenbaum creates ELIZA, a simple chatbot that mimicked a therapist using pattern matching. Users found it surprisingly engaging despite its simplicity.
1980s–90s: Statistical language models emerge, calculating probabilities of word sequences from text corpora. These power early spell-checkers and speech recognition.
2013: Word2Vec shows that neural networks can learn meaningful word representations—words with similar meanings cluster together in mathematical space.
2017: Google researchers publish "Attention Is All You Need," introducing the **Transformer** architecture. This breakthrough enables models to process entire sequences in parallel and track long-range dependencies.
2018–2020: OpenAI releases GPT (2018), GPT-2 (2019), and GPT-3 (2020), each dramatically larger than the last. GPT-3's 175 billion parameters demonstrate "emergent" capabilities not present in smaller models.
2022: ChatGPT launches, making LLMs accessible to the public. The conversational interface transforms how people interact with AI.
2023–Present: Rapid proliferation—GPT-4, Claude, Gemini, Llama, Mistral. LLMs become embedded in search engines, productivity tools, and countless applications.

The key insight of modern LLMs is that *scale matters*. Larger models trained on more data don't just perform incrementally better—they exhibit qualitatively new capabilities. A model with 1 billion parameters might struggle with basic reasoning; one with 100 billion parameters might solve complex problems. This "scaling law" has driven the race to build ever-larger models.

2 How LLMs Work: The Core Ideas

2.1 It's All About Prediction

Strip away the hype, and an LLM does one thing: given a sequence of text, predict what comes next. If you input "The capital of France is," the model predicts that "Paris" is the most likely next word.



To generate longer text, the model predicts one word (actually one **token**—more on that shortly), appends it to the input, and predicts again. A response of 100 words requires 100 sequential predictions, each building on all previous text.

Conceptual Text Generation Loop

```
public String generateText(String prompt, int maxTokens) {
    String text = prompt;

    for (int i = 0; i < maxTokens; i++) {
        // Get probability distribution over all possible next tokens
        Map<String, Double> probs = model.predictNext(text);
```

```
            // Sample from the distribution (or pick the most likely)
            String nextToken = sampleFrom(probs);

            // Append and continue
            text = text + nextToken;

            if (nextToken.equals("<END>")) break;
        }

        return text;
    }
```

This simple loop—predict, append, repeat—generates everything from haikus to legal briefs. The magic lies not in the loop but in the `predictNext` function: the neural network that has learned, from trillions of words, what text typically follows what.

2.2 Tokens: The Atoms of Language

LLMs don't process words directly—they process **tokens**, which are chunks of text that might be words, parts of words, or punctuation. The word "understanding" might be split into "under" + "stand" + "ing." Common words like "the" are single tokens; rare words get broken into pieces.

Text: "Tokenization is fascinating!"

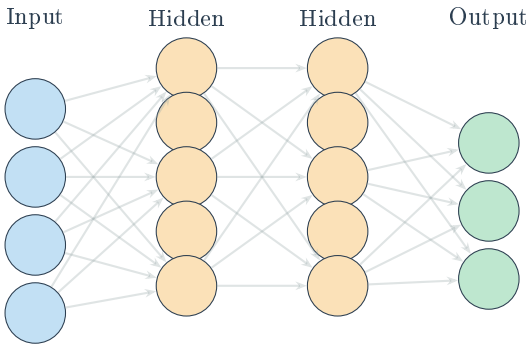
Tokens: `Token` `ization` `_is` `_fasci` `nating` `!`

(6 tokens — underscores represent spaces)

Tokenization matters because it determines what the model can "see." A model with a vocabulary of 50,000 tokens can represent any text as a sequence of token IDs—integers that index into the vocabulary. The model processes these numbers, not letters.

2.3 Neural Networks: Learning Patterns

How does the model learn to predict? Through a **neural network**—a mathematical structure loosely inspired by the brain. A neural network consists of layers of "neurons" (really just numbers) connected by "weights" (more numbers). Input flows through the network, transformed at each layer, until output emerges.



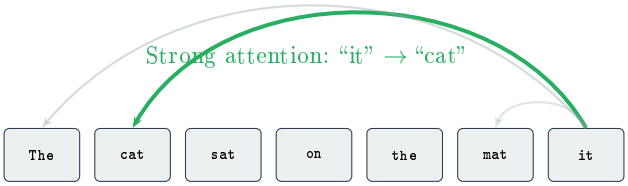
Each connection has a learnable weight

During **training**, the network sees billions of examples: given this text, the next token was X. If the network predicts wrong, the **loss** (error) is calculated, and the weights are adjusted slightly to make that prediction more accurate next time. Repeat this billions of times, and patterns emerge: the network learns grammar, facts, reasoning styles—everything implicit in human text.

2.4 The Transformer Architecture

Modern LLMs use a specific neural network design called the **Transformer**. Its key innovation is the **attention mechanism**, which allows every token to "attend to" every other token in the sequence.

When processing "The cat sat on the mat because it was tired," the model must understand that "it" refers to "cat," not "mat." Attention lets the model learn these connections: when processing "it," the model can look back at all previous words and weight their relevance.



Transformers process all tokens in parallel (unlike older sequential models), making them much faster to train. Multiple "attention heads" capture different types of relationships—some might track grammar, others semantics, others long-range dependencies. Stacking many transformer layers (GPT-4 reportedly has over 100) creates a deep network that extracts increasingly abstract patterns.

2.5 What the Model "Knows"

An LLM doesn't store facts in a database—it encodes patterns in billions of numerical weights. When you ask "What is the capital of France?," no lookup occurs. Instead, the network's weights have been shaped by training so that, given this input pattern, the output pattern corresponds to "Paris."

This is why LLMs can be confidently wrong (**hallucinate**). They're not retrieving verified facts; they're generating text that *statistically resembles* correct answers. Usually the patterns align with truth—but not always.

3 Training: Where Knowledge Comes From

3.1 The Training Data

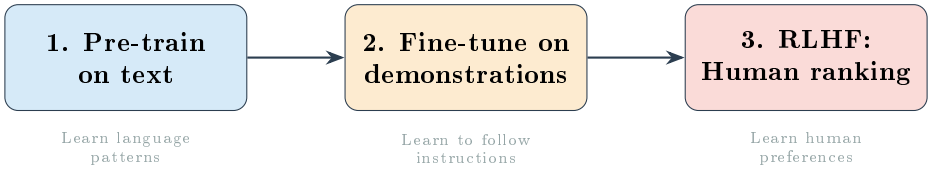
An LLM's capabilities come from its training data—the text it learned from. GPT-3 trained on roughly 500 billion tokens from sources including Common Crawl (web pages), books, Wikipedia, and code repositories. Later models use even more data.

This training data shapes everything about the model:

- **Knowledge**: The model only "knows" what appeared in training data (up to its cutoff date).
- **Style**: It mimics the writing styles present in training.
- **Biases**: Biases in human text become biases in the model.
- **Languages**: It performs best on languages well-represented in training.

3.2 RLHF: Teaching Models to Be Helpful

Raw language models trained only to predict text can be erratic—they might generate toxic content, refuse to help, or produce unhelpful responses. To make models like ChatGPT useful and safe, companies use **Reinforcement Learning from Human Feedback** (RLHF).



In RLHF, human raters compare model outputs and indicate which is better. The model learns to produce responses that humans prefer—more helpful, more accurate, less harmful. This process is expensive and labor-intensive, relying on thousands of human hours.

4 Ethical Issues in LLM Development and Deployment

The power of LLMs brings profound ethical challenges. These aren't hypothetical future concerns—they're happening now.

4.1 Bias and Discrimination

LLMs learn from human-generated text, which contains human biases. Studies have found that LLMs associate certain professions with genders ("nurse" with female, "engineer" with male), generate more negative sentiment for certain ethnic names, and reproduce harmful stereotypes.

Case: Bias in Image Generation

When text-to-image systems (built on similar technology) were asked to generate images of "a CEO," they overwhelmingly produced images of white men. Asked for "a nurse," they produced women. These systems learned to reproduce—and arguably amplify—the biases present in their training data, including photo captions and image descriptions from the internet.

The challenge is that bias is embedded in the training data itself. Debiasing techniques exist but are imperfect. And "removing bias" requires defining what unbiased means—a contested question.

4.2 Hallucination and Misinformation

LLMs confidently generate false information. They'll invent scientific citations, describe historical events that never happened, and provide plausible-sounding but incorrect technical advice. This isn't malfunction—it's inherent to how they work. They generate *probable* text, not *verified* text.

In high-stakes domains—medicine, law, journalism—hallucinations can cause real harm. A lawyer who submitted AI-generated legal briefs discovered the cited cases didn't exist. Medical chatbots have provided dangerous health advice. The fluent, confident style of LLM output makes hallucinations harder to detect.

4.3 The Labor Behind the AI

Creating "aligned" AI requires enormous human labor, often invisible and poorly compensated:

The Hidden Workforce

Training data must be cleaned and labeled. RLHF requires thousands of human comparisons. Content moderation—filtering toxic outputs—exposes workers to disturbing material. Much of this work is outsourced to workers in Kenya, the Philippines, and other countries, sometimes paying less than $2 per hour. Investigations revealed that workers screening training data for OpenAI faced traumatic content including violence and abuse.

The gleaming AI interfaces obscure this global supply chain of human labor.

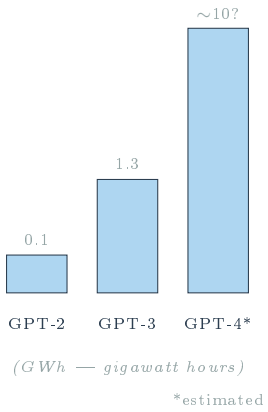
4.4 Copyright and Training Data

LLMs are trained on text scraped from the internet, much of it copyrighted: news articles, books, academic papers, personal blogs, code repositories. Authors and publishers have filed lawsuits arguing this constitutes copyright infringement.

The questions are genuinely difficult: Is training on copyrighted text "fair use"? Can a model that has "learned" from a book reproduce substantial portions? If an LLM writes in a distinctive author's style (having trained on their work), who owns that output? These cases are working through courts worldwide.

4.5 Environmental Impact

Training large models consumes enormous computational resources—and electricity. Training GPT-3 reportedly consumed enough energy to power an average American home for over 100 years. As models grow larger and companies race to train new versions, the environmental footprint grows.



Running inference (generating responses) also consumes energy. Every ChatGPT query uses roughly 10 times the energy of a Google search. Multiply by billions of queries, and the climate impact is significant.

4.6 Concentration of Power

Building frontier LLMs requires billions of dollars in compute, access to massive datasets, and teams of hundreds of researchers. Only a handful of companies—OpenAI, Google, Anthropic, Meta, a few others—can compete at the frontier. This concentrates enormous power over a transformative technology in few hands.

These companies decide what the models can and cannot do, what values they embody, and who gets access. Decisions made in San Francisco boardrooms shape how billions of people interact with information.

4.7 Job Displacement

LLMs can perform tasks previously requiring human expertise: writing, coding, customer service, legal research, medical triage. While technology has always displaced jobs while creating new ones, the breadth of LLM capabilities raises questions about whether this time is different.

Some jobs may transform rather than disappear—programmers using AI assistants rather than being replaced. But for workers whose skills closely match what LLMs do cheaply, the transition may be painful.

4.8 Manipulation and Misuse

LLMs can generate persuasive text at scale: personalized phishing emails, targeted political propaganda, fake reviews, synthetic social media personas. The same capabilities that make LLMs useful make them dangerous tools for manipulation.

Detecting AI-generated text is difficult and getting harder as models improve. A world where any text might be AI-generated undermines trust in written communication itself.

5 Open Questions and Active Debates

The previous sections described what we know with reasonable confidence about how LLMs work. But many of the most important questions remain genuinely unsettled—debated actively by philoso-

phers, computer scientists, and cognitive scientists. Rather than telling you what to think, this section presents the debates themselves.

5.1 From Chatbots to Agents

LLMs began as text generators, but they're increasingly being incorporated into **agents**—systems that can take actions in the world. An LLM agent might browse the web, write and execute code, send emails, book appointments, or control other software. Projects like AutoGPT, LangChain agents, and Claude's computer use represent early steps toward AI systems that don't just respond but *act*.

This raises new questions. A chatbot that gives bad advice is problematic; an agent that takes bad actions could cause direct harm. How much autonomy should AI agents have? What safeguards are needed? How do we maintain meaningful human oversight as agents become more capable? These are active areas of research with no consensus answers.

5.2 Understanding, Intelligence, and Agency

Perhaps the deepest debate concerns what LLMs *are*—or aren't. Consider several positions:

The Debate Over Understanding

The skeptical view: LLMs are "stochastic parrots" (a term from researchers Emily Bender and Timnit Gebru)—sophisticated pattern-matchers that produce plausible text without any genuine understanding. They manipulate symbols without grasping meaning, like someone following Chinese translation rules without knowing Chinese (philosopher John Searle's famous "Chinese Room" argument).

The moderate view: LLMs have developed genuine but limited competencies. They've learned something real about language, reasoning, and world knowledge—but this falls short of human understanding. They might have "functional understanding" for specific tasks without general comprehension.

The expansive view: Understanding exists on a spectrum, not as binary. LLMs exhibit behaviors that, in humans, we'd attribute to understanding. Perhaps our intuitions about what "true understanding" requires are too restrictive, or rooted in human chauvinism. If a system can reason, explain, and apply knowledge flexibly, what more should we demand?

There is no scientific consensus. The question may be philosophical rather than purely empirical—or it may require new concepts that don't map onto existing categories.

Similar debates surround **intelligence** and **agency**. Some argue LLMs demonstrate a form of intelligence; others insist intelligence requires embodiment, intrinsic goals, or consciousness that LLMs lack. Some see LLM-based agents as exhibiting genuine agency; others argue that without autonomous goal-formation, they remain sophisticated tools. These aren't merely semantic disputes—they affect how we should regulate AI, assign responsibility, and think about the future.

5.3 Authorship and Creativity

When an LLM writes a poem, who—if anyone—is the author? Consider the candidates:

- **The human who prompted it?** They provided creative direction, but perhaps not the execution.

- **The LLM itself?** It generated the specific words, but can something without intentions or experiences be an "author"?
- **The developers who built it?** They created the system, but didn't write this specific poem.
- **The authors of the training data?** The LLM's capabilities derive from their work, often used without consent.
- **No one / everyone?** Perhaps AI-generated content challenges our concept of authorship itself.

This matters beyond philosophy. Copyright law traditionally requires human authorship. Academic integrity policies assume individual responsibility for work. Creative communities debate whether AI art is "real" art. How we answer the authorship question shapes law, norms, and creative practice.

5.4 Near-Term vs. Long-Term Risks

The AI safety community itself is divided about which risks deserve priority:

Two Views on AI Risk

Near-term focus: The most pressing risks are happening now: bias and discrimination in deployed systems, job displacement, misinformation at scale, concentration of power, erosion of privacy, and environmental costs. These harms are concrete, measurable, and affect real people today. Speculating about superintelligence distracts from urgent problems we can actually address.

Long-term focus: If AI systems become sufficiently advanced, the stakes become existential. A misaligned superintelligent AI could pose catastrophic risks—not from malevolence, but from pursuing goals that don't properly account for human values. Even a small probability of such outcomes warrants serious attention given the magnitude of potential harm.

The tension: Resources and attention are finite. Focusing on long-term risks might neglect current harms. Focusing only on current harms might leave us unprepared for transformative AI. Some researchers argue these concerns are complementary; others see them as competing for limited resources.

Where you stand on this debate shapes views on AI governance, research priorities, and development pace. Reasonable, informed people disagree profoundly—and the stakes couldn't be higher.

6 Navigating Uncertainty

Given genuine uncertainty about what LLMs are and what risks they pose, how should we approach them? Perhaps with epistemic humility:

Verify important claims. Whatever LLMs are, they can produce false statements. For anything consequential—facts, citations, medical or legal information—verify with authoritative sources.

Hold conclusions lightly. We don't fully understand these systems. Be wary of confident pronouncements—including your own—about what they can or cannot do, understand or fail to understand.

Disclose AI involvement. However we resolve questions of authorship, transparency matters. When AI contributes to your work, acknowledge it.

Consider broader impacts. Individual uses aggregate into societal effects. Training data, labor practices, environmental costs, and automation effects connect your choices to larger systems.

Engage with the debates. The questions in this case study don't have settled answers. Form your own views through reflection and discussion—but remain open to evidence and argument. The conversation is just beginning.

Discussion Questions

Discussion Questions

1. **The Understanding Debate:** You've seen three positions on whether LLMs "understand": skeptical (stochastic parrots), moderate (limited functional understanding), and expansive (understanding on a spectrum). Which view do you find most compelling, and why? What evidence or argument might change your mind?

2. **Agents and Autonomy:** As LLMs are incorporated into agents that can take real-world actions, what level of autonomy is appropriate? Should an AI agent be able to send emails on your behalf? Make purchases? What safeguards would you want, and who should design them?

3. **Authorship and Creativity:** If you use an LLM to help write an essay, who is the author? Does it matter whether the LLM wrote 10% or 90%? How should academic integrity policies, copyright law, and creative norms adapt to AI-assisted work?

4. **Risk Priorities:** The AI safety community debates whether to prioritize near-term harms (bias, job loss, misinformation) or long-term risks (misaligned superintelligence). Where do you think attention and resources should focus? Can both be addressed simultaneously, or must we choose?

5. **Your Own Position:** After reading this case study, what's your overall view of LLMs? Are they powerful tools, emerging minds, dangerous technologies, overhyped products, or something else entirely? What uncertainties remain in your thinking?

Key Terms

Glossary of Key Terms

Agent	An AI system that can take actions in the world (browsing, coding, sending messages), not just generate text.
Agency	The capacity to act autonomously toward goals—whether LLMs or LLM-based systems possess genuine agency is debated.
Attention Mechanism	A technique allowing neural networks to weigh the relevance of different parts of input when processing each token.
Embedding	A vector of numbers representing a token's meaning, learned during

	training to capture semantic relationships.
Hallucination	When an LLM generates plausible-sounding but factually false information—an inherent risk of prediction-based systems.
Intelligence	The capacity for reasoning, learning, and problem-solving—whether LLMs exhibit genuine intelligence remains contested.
Large Language Model	A neural network with billions of parameters trained on massive text data to predict and generate language.
Neural Network	A computational model with layers of connected nodes whose connection strengths (weights) are learned from data.
RLHF	Reinforcement Learning from Human Feedback; training models to produce outputs that humans rate as preferable.
Stochastic Parrot	A critical term suggesting LLMs merely produce statistically likely text without genuine understanding.
Token	A unit of text (word, subword, or punctuation) that serves as input and output for language models.
Training	The process of adjusting neural network weights by exposure to billions of examples to minimize prediction error.
Transformer	The neural network architecture underlying modern LLMs, using attention mechanisms and parallel processing.
Understanding	Grasping meaning and implications—whether LLMs possess any form of understanding is actively debated.