Economics of AI

Companion Reading List & Resource Guide

Ashish Kulkarni | Takshashila Institution | GCPP Program | February 2026

Before You Begin: Make This Document Yours

This reading list accompanies the 90-minute lecture on the Economics of AI. It is organized to mirror the presentation's upstream/midstream/downstream structure. Each section includes key concepts, recommended readings, and links for deeper exploration.

But here's the thing: a reading list someone else made for you is useful. A reading list you build yourself is transformative. This document is designed to help you do both.

How to use this guide: Each section has a "Start here" recommendation—if you read nothing else, read that. You'll also find **Deep Research Prompts** throughout the document. These are ready-to-use prompts you can paste into any LLM (Claude, ChatGPT, Gemini, or any other) to generate your own in-depth research reports on each topic. Use the "Deep Research" or equivalent feature in your LLM of choice for the best results, but even a regular conversation will give you a solid foundation.

Create and share your own reading list. As you work through this document, you'll generate research reports, discover new sources, and develop your own perspective on the economics of AI. We'd like you to compile your best findings into your own annotated reading list and **share it in the GCPP shared Google Drive folder** so that every student has access to every other student's reading list. Thirty students generating thirty reading lists means everyone leaves this course with a richer, more diverse set of resources than any one person could assemble alone. The shared folder is your cohort's collective knowledge base—contribute to it generously.

UPSTREAM: The Cost Structure of Producing AI

The upstream section examines the inputs required to produce AI: silicon and semiconductor fabrication, data center infrastructure, training data, compute costs (training vs. inference), energy, and human talent. The central insight is that AI production is dominated by massive, irreversible fixed costs—which is why so few firms can afford to compete at the frontier.

Silicon & Infrastructure (Modules 1–2)

Key concepts: Moore's Law (transistor density doubles every ~2 years) vs. Rock's Law (fab costs double every ~4 years). The semiconductor industry has consolidated to just three firms capable of cutting-edge fabrication: TSMC, Samsung, and Intel. A single fab now costs $20B+. AI training clusters ("gigaclusters") require thousands of specialized GPUs, power substations, cooling infrastructure, and high-bandwidth interconnects.

Readings

⭐ *Start here: Generate your own deep research report on AI gigaclusters.*

> 🔎 **Deep Research Prompt:** *"I'm studying the economics of AI infrastructure. Write a comprehensive research report on AI gigaclusters: what they are, what hardware they require (GPUs, networking, power), how much they cost to build and operate, who is building them, and what the economic implications are of concentrating so much compute in single facilities. Include specific examples (Meta's Llama 3 training infrastructure, Microsoft's clusters, Google's TPU pods) and discuss how gigacluster economics create entry barriers in the AI industry."*
>
> 🔎 **Deep Research Prompt:** *"Write a research report on the economics of AI chips and semiconductor geopolitics. Cover the foundry model (TSMC, Samsung, Intel), the cost structure of chip fabrication, Nvidia's dominance in AI GPUs vs. custom silicon efforts (Google TPUs, Amazon Trainium), and the geopolitical dimensions: Taiwan's centrality, US export controls on China, and efforts to reshore semiconductor manufacturing. How do chip economics shape who can and cannot compete in AI?"*

Chris Miller, *Chip War: The Fight for the World's Most Critical Technology* (2022). The definitive popular account of semiconductor geopolitics. Essential background for understanding why chips matter for AI economics.

For a quick overview of gigacluster costs, see reporting on Meta's Llama 3 training infrastructure ($720M in hardware alone for 24,000 H100 GPUs) and Sam Altman's vision of 1GW clusters coming online weekly by 2030.

Data Acquisition & Preparation (Module 3)

Key concepts: Training data has transformed from an afterthought to a multi-billion dollar market. The shift from free web scraping to expensive licensing (driven by copyright settlements like Anthropic's $1.5B settlement in 2025) has fundamentally altered the economics. High-quality public text data is projected to be exhausted between 2026–2032. Synthetic data offers cost savings (60–80% cheaper) but introduces "model collapse" risks. Data preparation constitutes 15–25% of total AI development costs.

Readings

⭐ *Start here: Generate your own deep research report on the data economy behind AI.*

> 🔎 **Deep Research Prompt:** *"Write a comprehensive research report on AI training data economics. Cover: the evolution of data sourcing (web scraping to licensing), the cost structure of data acquisition and preparation, the data licensing market (major deals, pricing models, copyright implications), quality vs. quantity tradeoffs, the economics of RLHF and human feedback, synthetic data (costs, benefits, model collapse risks), and projections for when high-quality public text data will be exhausted. What are the economic implications of data becoming scarce?"*

Villalobos et al., "Will We Run Out of Data? Limits of LLM Scaling Based on Human-Generated Data," arXiv:2211.04325 (2022). The key paper on data exhaustion timelines.

Microsoft's Phi model series papers (available on arXiv) demonstrate 1,000–10,000x cost improvements through data curation rather than scale—a powerful counterpoint to the "bigger is better" narrative.

Training vs. Inference & Cost Decay (Modules 4–5)

Key concepts: Training is a one-time CapEx (building the factory); inference is ongoing OpEx (running it). AWS estimates inference accounts for 90% of total model cost. GPT-3.5-level inference costs dropped 280x between 2022–2024 due to hardware optimization, quantization, distillation, and competition. But the Jevons Paradox applies: cheaper AI drives more usage, so total energy and compute consumption rises even as per-query costs fall. The "DeepSeek moment" (early 2025) proved that efficiency innovations can rival raw scale.

Readings

⭐ *Start here: Generate your own report on the training vs. inference cost divide.*

> 🔎 **Deep Research Prompt:** *"Write a research report comparing AI training costs vs. inference costs for major language models (GPT-3, GPT-4, Claude, Llama). Cover: the economics of training runs (hardware, time, energy, total cost), inference cost structure (per-token pricing, optimization techniques like quantization, distillation, speculative decoding), and how the balance between training and inference spending has shifted over time. Include the AWS estimate that inference is 90% of total cost and discuss implications for business models."*

> 🔎 **Deep Research Prompt:** *"Write a research report on the cost decay phenomenon in AI. How have the costs of AI inference fallen over time? Apply Wright's Law and learning curves to AI cost data. Cover the shift from process innovation to product innovation, the role of competition (especially DeepSeek) in driving costs down, dynamic pricing strategies under falling marginal costs, and the Jevons Paradox as applied to AI compute. What does the history of cost decay in semiconductors, solar panels, and genomics tell us about AI's trajectory?"*

Hestness et al., "Deep Learning Scaling Is Predictable, Empirically," arXiv:1712.00409 (2017). The foundational paper on scaling laws.

Ho et al., "Algorithmic Progress in Language Models," NeurIPS 2024. Quantifies how algorithmic improvements reduce compute requirements over time.

Energy, Environment & Talent (Modules 6–8)

Key concepts: AI is "recoupling" economic growth with electricity demand after decades of decoupling. US data centers may consume 12% of total electricity by 2028. "Energy gentrification" describes how data centers compete with local communities for finite grid capacity. Water consumption is the "secret footprint"—a single large data center uses 5M gallons/day. Talent economics are extreme: top researchers earn $1–10M+, and the data labeling workforce (concentrated in Kenya, Philippines, India) represents an invisible but critical labor force.

Readings

⭐ *Start here: Generate your own report on AI's energy and environmental footprint.*

> 🔎 **Deep Research Prompt:** *"Write a research report on the energy and environmental economics of AI. Cover: unit economics of energy consumption (training vs. inference), Power Purchase Agreements (PPAs) and how AI companies secure energy, geographic constraints on data center placement, water consumption, the Jevons Paradox applied to AI energy use, and the concept of 'energy gentrification.' Include projections for US data center electricity consumption and discuss whether AI will recouple economic growth with energy demand."*

> 🔎 **Deep Research Prompt:** *"Write a research report on the talent and labor economics of AI. Cover three layers: (1) elite AI researchers (compensation, acqui-hires, geographic concentration), (2) the data labeling workforce (Kenya, Philippines, India—working conditions, pay, geographic arbitrage), and (3) infrastructure workers (data center construction and operations). How do talent costs compare to compute costs? What are the implications of the extreme compensation gap between researchers and data labelers?"*

Aljbour, Wilson & Patel, "Powering Intelligence: Analyzing AI and Data Center Energy Consumption," EPRI White Paper 3002028905 (2024).

> 🔎 **Deep Research Prompt:** *"Write a research report on the economics of AI regulatory compliance. How do regulations (EU AI Act, US executive orders, China's approach) function as entry barriers? What are the compliance cost structures for frontier labs vs. startups vs. open-source projects? Compare the US, EU, and China regulatory approaches and their economic implications for market structure and innovation."*

MIDSTREAM: Market Structure & Competitive Dynamics

The midstream section examines who produces AI and why the market is structured the way it is. It covers the oligopolistic market structure, entry barriers, the role of open source, vertical integration, supply chain dynamics, and the geopolitical fragmentation of the AI industry.

The Two-Tier Oligopoly (Modules 9–10)

Key concepts: The frontier AI market is a two-tier oligopoly. Tier 1 (Anthropic, OpenAI, Google DeepMind) competes at the cutting edge with proprietary models, massive capital, and differentiated strategies. Tier 2 (xAI, Mistral, others) differentiates rather than competing head-on. Entry barriers are self-reinforcing: capital buys talent, talent creates data moats, data moats require scale, and scale requires capital. As of February 2026, the latest frontier releases include Claude Opus 4.6 (1M token context, Agent Teams) and GPT-5.3 Codex (self-debugging training runs).

Readings

⭐ *Start here: Generate your own market structure analysis of the AI industry.*

> 🔎 **Deep Research Prompt:** *"Write a research report on the market structure of the frontier AI industry. Apply industrial organization frameworks (Cournot and Bertrand oligopoly models, quality ladders, HHI and Lerner index) to analyze competition among Anthropic, OpenAI, Google DeepMind, Meta, and others. Who are the Tier 1 and Tier 2 players? How do they differentiate? What market concentration measures tell us about competitive dynamics? Include the latest model releases and competitive moves."*

> 🔎 **Deep Research Prompt:** *"Write a research report on entry barriers in the AI industry. Analyze both structural barriers (capital requirements, data moats, talent scarcity, compute access) and strategic barriers (limit pricing, exclusive partnerships, vertical integration). How do these barriers reinforce each other? Apply concepts from industrial organization: experience economies, minimum efficient scale, and contestable markets theory. Is the AI oligopoly stable or vulnerable to disruption?"*

For the latest competitive landscape, see the announcement posts for Claude Opus 4.6 (Anthropic blog, Feb 5, 2026) and GPT-5.3 Codex (OpenAI blog, Feb 5, 2026)—both released on the same day.

Open Source & Competitive Strategies (Modules 11–12)

Key concepts: Meta's Llama strategy follows the logic of "commoditize the complement": give away the model to profit from the ecosystem (advertising, engagement). Open source creates a price ceiling for closed models and narrows the quality gap. The sustainability question is whether open source can match frontier capabilities without billions in capital—the answer appears to be "only if backed by a tech giant with different incentives." The analogy is Google Maps: free GPS navigation destroyed standalone GPS devices because Google profits from knowing where you go, not from selling maps.

Readings

⭐ **Start here:** Joel Spolsky's classic essay "Strategy Letter V" (2002) on "commoditize your complement." The single best framework for understanding Meta's Llama strategy. Available at joelonsoftware.com.

The DeepSeek technical report (available on arXiv) demonstrates how efficiency-focused approaches can achieve competitive performance at a fraction of the cost of frontier labs, challenging the assumption that scale is everything.

Supply Chain, Vertical Integration & Geopolitics (Modules 13–18)

Key concepts: Nvidia holds 80–90% of the AI GPU market, with CUDA creating powerful software lock-in. Vertical integration is accelerating (Google makes chips + models + apps; Microsoft-OpenAI and Anthropic-Amazon represent quasi-integration through strategic partnerships). US-China chip export controls are creating parallel technology ecosystems. India's position is primarily in the services layer (data labeling, RLHF workforce, IT services integration) rather than frontier model production.

Readings

⭐ **Start here:** Ben Thompson, Stratechery—his ongoing analysis of AI market structure, vertical integration, and aggregation theory applied to AI is the best running commentary available. Key posts on Nvidia's moat, the Microsoft-OpenAI relationship, and platform dynamics.

For India's AI policy position, Takshashila Institution's own research papers on India's digital economy and AI governance are highly relevant context.

DOWNSTREAM: Markets, Business Models & Economic Impacts

The downstream section examines who benefits from AI and how. It covers pricing strategies, the emerging agentic economy, labor market transformation, productivity measurement, consumer surplus, and the distribution of gains from AI.

Pricing & Business Models (Modules 19–27)

Key concepts: Consumer pricing has converged at $20/month (Plus) and $200/month (Pro) across both OpenAI and Anthropic—raising questions about tacit coordination. Revenue comes from three layers: consumer subscriptions, API per-token pricing, and enterprise contracts ($100K–millions/year). The consumer surplus gap is enormous: users paying $20/month may extract $500/month in value. This suggests AI may be the most underpriced technology in history relative to the value it creates.

Readings

⭐ **Start here:** The current pricing pages of Claude (anthropic.com) and ChatGPT (openai.com) are themselves informative documents—study the tier structure, feature differentiation, and what's included at each level. Compare them side by side.

For consumer surplus estimation and the economics of digital goods, Hal Varian's work on information goods pricing remains foundational. His textbook *Information Rules* (co-authored with Carl Shapiro) is dated but the core economics are directly applicable.

The Agentic Economy (Emerging, 2025–2026)

Key concepts: The shift from "AI as tool" to "AI as autonomous worker." The agentic AI market is valued at $4.5B (2025) and projected to reach $98B by 2033. Key developments include Claude Code (terminal-native coding agent using decide→do→verify loops), GPT-5.3 Codex (which debugged its own training run), and OpenClaw/Moltbook (a personal AI assistant with 150K+ GitHub stars, and a social network for AI agents with 1.7M agent accounts). The Model Context Protocol (MCP), donated to the Linux Foundation with 97M monthly SDK downloads, is standardizing how agents connect to tools—think of it as HTTP for agent-to-service communication.

Readings

⭐ **Start here:** Anthropic's Claude Code documentation (code.claude.com/docs) provides the clearest picture of what agentic coding looks like in practice—how agents gather evidence iteratively, work with git, and create specialized subagents.

MIT Technology Review, "Moltbook Was Peak AI Theater" (Feb 6, 2026). A critical assessment of the social-network-for-agents concept—useful for understanding both the hype and the genuine signal about where agentic AI is heading.

Deloitte, "Agentic AI Strategy" (2026 Tech Trends). Covers enterprise adoption patterns, the shift from single agents to orchestrated multi-agent teams, and the 40% enterprise embedding prediction by end of 2026.

For the Model Context Protocol, the MCP specification documentation and the Linux Foundation announcement provide the technical foundation.

Labor Markets, Productivity & Inequality (Modules 28–38)

Key concepts: Solow's Paradox (computers everywhere except in the productivity statistics) is repeating with AI—or is it? The Acemoglu-Restrepo task-based framework is essential: AI automates tasks, not entire jobs. Jobs are bundles of tasks; some get automated, others get complemented. The implementation J-curve means productivity dips before rising. Historical precedent suggests electricity took 40 years, computers 20–30 years, and the internet 10–15 years to show up in productivity statistics—AI may be faster. Five dimensions of inequality matter: capital vs. labor, AI-adopters vs. laggards, high-skill vs. low-skill, producers vs. consumers (geographically), and present vs. future.

Readings

⭐ **Start here:** Daron Acemoglu & Pascual Restrepo, "Tasks, Automation, and the Rise in U.S. Wage Inequality," *Econometrica* (2022). The foundational paper for the task-based framework. Technical but the core argument is accessible.

Erik Brynjolfsson et al., studies on GitHub Copilot productivity (2023). One of the best empirical studies on AI's actual productivity effects, showing differential impacts across skill levels.

Ouyang et al., "Training Language Models to Follow Instructions with Human Feedback," NeurIPS 2022. The foundational RLHF paper—relevant for understanding both the technical process and the invisible labor behind AI alignment.

David Autor's work on job polarization and the "hollowing out" of middle-skill work provides essential context for understanding how AI differs from (and resembles) previous waves of automation.

Shulman & Bostrom, "How Hard Is Artificial Intelligence?" Journal of Consciousness Studies (2012). An older but thought-provoking paper on evolutionary arguments about AI difficulty—useful for framing long-term questions about comparative advantage between humans and AI.

Further Exploration: Cross-Cutting Themes

These resources cut across the upstream/midstream/downstream structure and offer broader perspectives on the economics of AI.

On the Economics of AI More Broadly

Marginal Revolution University (MRU) videos on microeconomic theory provide excellent background for the economic concepts used throughout this course (market structure, externalities, consumer surplus, comparative advantage). Free at mru.org.

Vafa et al., “Evaluating the World Model Implicit in a Generative Model,” NeurIPS 2024. A fascinating paper on what LLMs actually “know”—relevant for understanding the nature of the product these firms are producing.

Goldstein & Levinstein, “Does ChatGPT Have a Mind?” arXiv (2024). Provocative philosophical analysis with implications for how we value and price AI capabilities.

On India & Emerging Markets

Takshashila Institution’s own policy research on India’s digital economy, AI governance, and technology policy.

The data labeling industry in India (Scale AI, Surge AI operations) is an important case study in geographic arbitrage and global labor economics. Investigative reporting by Time, The Guardian, and others on data labeling working conditions provides the human dimension.

Ongoing Sources Worth Following

- Stratechery (Ben Thompson)—the best ongoing analysis of AI business strategy and market structure.
- Epoch AI (epoch.ai)—rigorous, data-driven research on AI compute trends, data exhaustion, and scaling laws.
- The Information, Bloomberg, and TechCrunch—for the latest on AI company financials, deals, and market developments.
- NBER Working Papers on AI economics—search for papers by Acemoglu, Brynjolfsson, Agrawal, and Autor for the latest academic research.
- IEA (International Energy Agency) annual reports on data center energy consumption and projections.

A Note on This Document

This is intended as a living document. The economics of AI is evolving rapidly—the competitive landscape, pricing models, regulatory environment, and technological capabilities are all shifting in real time. The upstream/midstream/downstream framework, however, is durable. New developments will slot into this structure even as specific facts and figures become outdated.

The Deep Research Prompts throughout this document are designed to help you generate up-to-date reports whenever you need them. The readings may become dated; the prompts will always produce current analysis. Use both.

If you find interesting resources that fit within this framework, add them to your own reading list and share it with your cohort in the shared Google Drive folder. The best learning happens when it’s collaborative.

— Ashish Kulkarni, February 2026