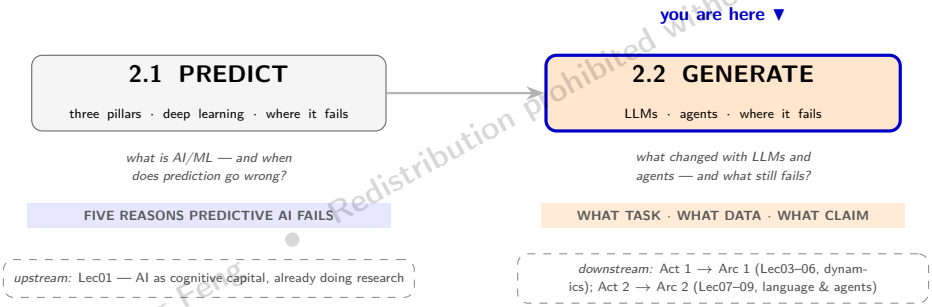
AI for Economic Research: Dynamic Models, Language, and Agents
Lecture 2.2: Language, Generation, and Agency

Zhigang Feng

2026-07-07 10:17:24

Lecture 2 in One Map

One lecture, two decks, one taxonomy: the predictive stack and the generative stack — what each is, how it is used in economics, and *where each fails*.



Act 2.2 (here) *builds the generative stack: the intelligence debate, NLP → LLMs → agents, how people actually use generative AI — then its own caution file (hallucination, deepfakes) and the practitioner's discipline: what task, what data, what claim.*

What Is Intelligence? The Cognitive Debate

What "understanding" means for a machine — induction, abduction, and the stochastic-parrot question

What is Intelligence?

- **Problem-Solving**: The capacity to identify problems and devise effective solutions.
- **Adaptability**: The ability to adjust behavior in response to changes in the environment.
- **Learning**: Acquiring new knowledge or skills through experience and education.
- **Reasoning**: Drawing conclusions based on available information, including logical reasoning and inference.
- **Classification and Generalization:**
  - Rosch (1975): people categorize objects by comparing them to an idealized prototype.
  - Chomsky (1965, *Aspects*): gives the modern generative formulation of universal grammar (reviving a much older idea) — all human languages share a common underlying structure.
  - *AI tries to learn these same cognitive capabilities from data.*

The Cognitive Debate: Stochastic Parrots vs. Reasoners

- **The Stochastic Parrot Hypothesis (Bender et al., 2021, FAccT '21):**
  - LLMs produce plausible text without communicative intent or world models.
  - Valid simulators of language — but not cognitive agents.
- **Chomsky's Critique:**
  - LLMs excel at **description and prediction** but fail at **explanation**.
  - True intelligence requires causal reasoning; LLMs rely on probabilistic correlation.

The dividing question: is fluent language *evidence of understanding*, or a convincing imitation of it? Economists face the same question whenever a model "fits" the data without explaining it.

The Reasoning Gap: Induction vs. Abduction

- **Where AI is strong — inductive reasoning (归纳推理):** generalizing from massive data distributions (more examples $\Rightarrow$ better).
- **Where humans are strong — abductive reasoning (溯因推理):** inferring the *best explanation* from limited evidence.
- **The benchmark gap (Chollet, 2019; ARC Prize 2024):** LLMs dominate semantic benchmarks (MMLU) yet struggle with the **Abstraction and Reasoning Corpus (ARC)**, which tests on-the-fly adaptation to *novel* abstract rules.

Bottom line for economists: lean on AI where inductive pattern recognition dominates; stay skeptical where novel *causal judgment* matters.

AI vs. Human Intelligence: Strengths and Limits

	Human Intelligence	AI (Current)
Pattern matching	Good	Superhuman
Novel causal reasoning	Strong	Weak
Explanation	Strong	Weak
Common sense	Strong	Improving
Replication cost	20+ years	Near-zero
Parallelism	One body	Millions of instances
Energy	$\sim$20W	Kilowatts (at scale)
Domain range	Broad	Narrow $\rightarrow$ General
Context memory	Long-term	Context window

From NLP to LLMs to Agents

The generative stack: systems that *generate* and *act*

Natural Language Processing: From Bag-of-Words to Transformers

- **Generation 1 — Bag-of-Words:**
  - Represent text as a vector of word counts, ignoring grammar and order
  - TF-IDF weights words by frequency in document vs. corpus
  - Application: basic document classification, keyword search
- **Generation 2 — Word Embeddings (Word2Vec, GloVe, 2013–2014):**
  - Map words to continuous vectors capturing semantic similarity
  - "king $-$ man $+$ woman $\approx$ queen"
  - Application: measuring sentiment, ideology shifts in text
- **Generation 3 — Transformers and LLMs (2017–present):**
  - Self-attention: weight each word by its relevance to every other word in the sequence
  - Pre-trained on internet-scale data; fine-tuned or prompted for specific tasks
  - Unified architecture handles text, code, images, time series

Large Language Models: Predicting the Next Word

- **One simple objective.** An LLM repeatedly predicts the *next token* given everything before it — and that single trick, at enormous scale, produces fluent, useful text:

$$P(w_1, w_2, \ldots, w_T) = \prod_{t=1}^{T} P(w_t \mid w_1, \ldots, w_{t-1})$$

(just next-word prediction, over and over.)

- **What it is built from** (mechanics in Lec. 7):
  - *Tokenization:* break text into subword pieces.
  - *Embeddings:* map tokens to vectors that capture meaning.
  - *Self-attention (the Transformer):* let each word weigh every other word, capturing long-range context.

LLMs: The Knowledge-Cutoff Problem

- An LLM is trained on a *static snapshot* of data: it does not "know" about events after its training cutoff, and cannot draw on a source it never saw.
- **The fix — Retrieval-Augmented Generation (RAG):** fetch relevant, up-to-date documents at query time and condition the model on them (Lec. 8).

Preview: full treatment in Lec. 7 (LLMs) and Lec. 8 (RAG) — how to feed *your* data and sources to an LLM.

Reasoning & Agentic AI (Preview)

- **The Reasoning & Agentic Era (2024–):**
  - *Reasoning models* (OpenAI o1/o3, Claude thinking, DeepSeek-R1) trade inference compute for chains of thought — a partial answer to Chollet's ARC critique (Chollet, 2019; ARC Prize 2024).
  - *Open-source frontier* (Llama 3+, DeepSeek, Qwen) closes the gap with proprietary models; lowers the cost of academic experimentation by orders of magnitude.
  - *Agentic AI:* models act autonomously, plan multi-step tasks, write and execute code, and drive research workflows.

- **Why preview it here?** Generation (this deck) is the substrate; *agency* is generation wired into a loop with tools, memory, and goals.

Full treatment in Lec. 9 (Agentic AI); for research workflows that string these together, see the case studies in Lec. 10.

Technique Taxonomy: The Five Families

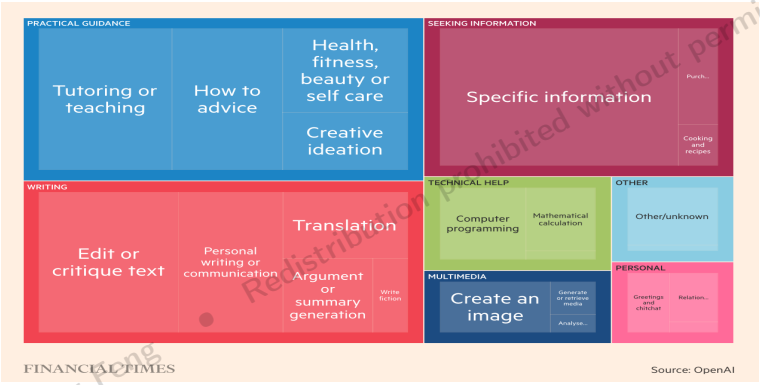
Family	What it is	Where in the course
Machine Learning	learn $f : x \to y$ from data (sup./unsup.)	Lec. 3
Deep Learning	neural nets as universal approximators	Lec. 3–4
Reinforcement Learning	sequential decisions = Bellman equation	Lec. 5–6
Large Language Models	transformers over token sequences; generation	Lec. 7–8
Agentic AI	LLMs + tools + planning loops	Lec. 9–10

One lineage: **ML** learns patterns → **DL** scales them → **RL** acts on them → **LLMs** generate language → **agents** act through language. The rest of the course walks this ladder.

How People Use Generative AI

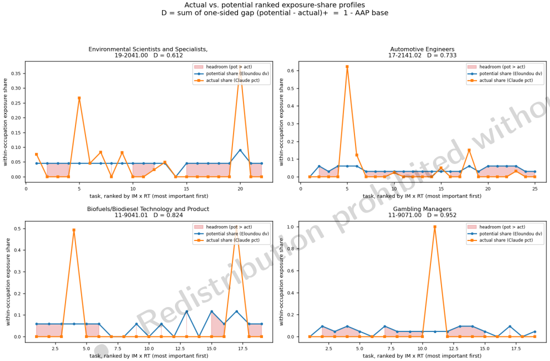
From everyday use to research use

How People Use Generative AI



Consumer usage is dominated by *generative* tasks — writing, coding, tutoring, and open-ended Q&A. But *who* uses it, and for *which* tasks, varies enormously across the economy — something we can now measure directly. →

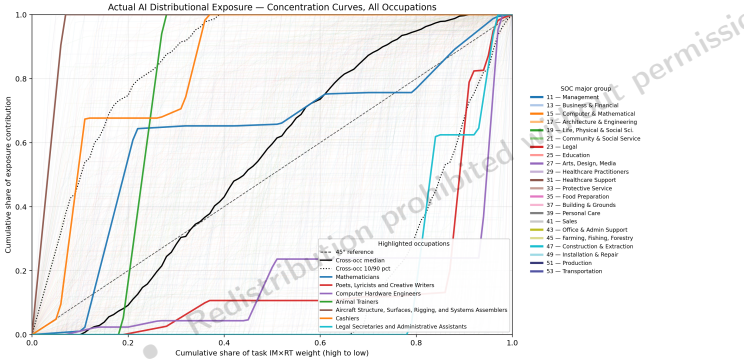
How AI Is *Actually* Used Within an Occupation



Each panel is one occupation; its tasks run along the x-axis, *most important first*. **Orange = where AI is actually used**; **blue = where it *could* be**. Actual use *spikes on one or two tasks* and leaves the rest untouched — and the pattern differs sharply by occupation.

*Source: author's analysis — Anthropic Economic Index (Claude usage) × O*NET importance × Eloundou et al. (2023).*

Zooming Out: AI-Usage Concentration Across 800+ Occupations



Each faint line is one occupation's *concentration curve* — how fast AI usage accumulates as tasks are added most-to least-important. *Above* the 45° line = usage on **core** tasks (Cashiers, Mathematicians); *below* = on **peripheral** tasks. AI's footprint is highly uneven.

*Source: author's analysis — Anthropic Economic Index (Claude usage), 833 O*NET occupations.*

Caution: Why Generative AI Still Fails

Hallucination and deepfakes: it *can*, but unreliably or harmfully

Hallucination: Automating Plausibility, Not Truth

- LLMs are trained to produce *plausible* text, not *true* statements: there is **no source of truth during training**. Fabrication is therefore *inherent* — not a bug a future patch removes.
- **Anchor (Mata v. Avianca, 2023):** a New York lawyer filed a ChatGPT-written brief that cited **entirely fabricated cases and invented judicial opinions**; asked whether they were real, the model insisted they were. The lawyer was sanctioned.
- **It is systematic, not a one-off:**
  - **CNET:** of 77 AI-written articles, **41 contained factual errors** — despite a "fact-checked" label.
  - *Defamation by chatbot:* models have fabricated news stories accusing real people of crimes they never committed.

Implication: generative AI is *unreliable for any high-stakes factual task* unless every claim is independently verified.

Source for this section's cases: Narayanan & Kapoor, *AI Snake Oil* (2024).

Deepfakes & Malicious Use: Capability Turned Against People

- The same generative capability produces realistic **voice, image, and video** from seconds of source material.
- **Fraud (2019):** a UK energy firm's managing director wired $\approx$ EUR 220,000 to a fake supplier after a **voice-cloned call** impersonating his boss — one of the first confirmed AI-enabled scams.
- **Non-consensual harm:** a single Telegram bot generated **100,000+ fabricated intimate images** of real women without their consent; the *large majority* of harmful deepfakes target women this way ("image-based abuse").
- **The "liar's dividend":** once *anything* could be fake, real evidence can be waved away as fake too — truth itself loses its anchor.

Implication: *capability* $\neq$ *safety* — a more capable generator is also a more capable weapon.

Open Weights and the Safety Frontier: The DeepSeek Case

A genuine capability leap — and a live case study in *capability* ≠ *safety*.

- **The upside.** DeepSeek's open-weight models (V3, R1; 2025) matched frontier reasoning at a fraction of the cost — exactly the "open-source frontier" that lowers the cost of academic experimentation.
- **The catch — weak guardrails.** Independent red teams bypassed R1's safety filters with ease: Cisco reported a *∼100% attack-success rate* on standard jailbreaks; others elicited malware, ransomware, and weapon-making instructions.
- **Open weights make safety *optional*.** Anyone can download the weights and *fine-tune the guardrails away*; safety that lives only inside the released model is not a real barrier ("illusory safety").
- **Governance fallout.** Data stored under PRC jurisdiction, an exposed user-log database, and topic censorship led *many governments* to ban it from official devices (Italy, Australia, South Korea, Taiwan, U.S. federal/state agencies).

For researchers: open models are invaluable for reproducibility and cost — but *openness shifts the safety burden onto you*, the deployer. Capability, cost, and safety are *separate* axes. *(Status as of 2026; red-team findings from Cisco, KELA, Palo Alto Unit 42.)*

The Practitioner's Discipline: What Task? What Data? What Claim?

- **Reliability $\neq$ capability.** The two failure modes differ in kind:
  - *Predictive failure* (T2.1): the model **can't** — the future isn't in the data (COMPAS, Fragile Families).
  - *Generative failure* (this deck): the model **can, but unreliably or harmfully** (hallucination, deepfakes).

- **The habit that guards against both** — before trusting any AI output, ask:
  1. **What task** is this, really? (a prediction? a generation? a decision?)
  2. **What data** did it learn from, and does my setting match?
  3. **What claim** am I making on its output, and how would I verify it?

AI is powerful *but bounded*. We return to this discipline in **Lec. 7 T3b** (LLM limitations) and turn it into a publication-grade **audit checklist in Lec. 9 T5**.

The Same Discipline, for Economists

The three questions, sharpened for research use:

1. **What task?** Is the AI output a *measurement* (a new variable), a *prediction*, or a *decision rule*? Each enters a paper differently — and a measured variable is usually the most defensible.

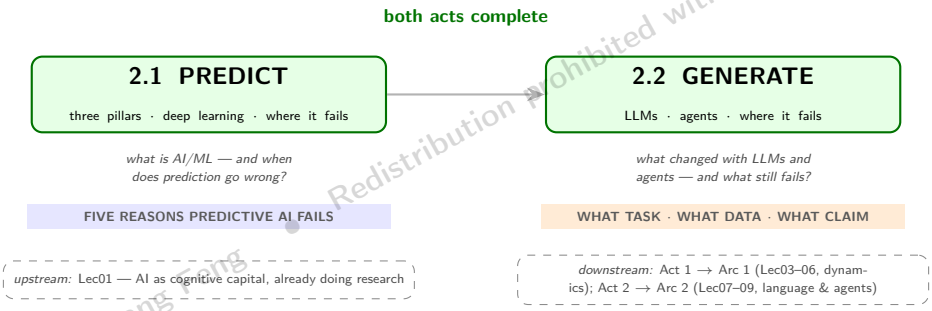
2. **What data?** Does the training distribution match *your* sample and period? Watch for *look-ahead / leakage* (a text model that already "knows" the outcome) and *regime shifts* — the Lucas-critique problem from T2.1.

3. **What claim?** If an AI-generated variable sits on the *right-hand side*, it carries *measurement error* — attenuation and generated-regressor problems. If it is the *outcome*, validate it against ground truth on a held-out, human-coded sample.

Rule of thumb: treat an AI output as an *estimated input*, never as ground truth — report the model, version, and prompt for *reproducibility*, and robustness-check the AI step as you would any first-stage estimate. (Toolkit: Lec. 7–9.)

Lecture 2, Complete: One Map, Two Acts

You now hold the taxonomy: the predictive stack and the generative stack, each with its powers and its documented failure modes.



The two acts now split into the course's two arcs: the predictive/RL stack drives Lec. 3–6 (dynamics); the generative stack drives Lec. 7–9 (language & agents). Ask always: what task, what data, what claim.

Preview: What Lec. 3 Delivers

- **Lec. 3 T1 — Why ML for macro:** unstructured data (text, satellite, panels), curse of dimensionality, pattern recognition as motivation.
- **Lec. 3 T2 — Neural networks:** architecture, activations (ReLU, sigmoid, tanh), universal approximation (Cybenko 1989, Hornik 1991).
- **Lec. 3 T3 — Training:** forward/back propagation, gradient descent + AdamW, mini-batches, regularization (dropout, weight decay).
- **Lec. 3 T4 — Specialized architectures:** ICNN, monotone NN, PICNN with macro shape restrictions.

Lec. 3 is the toolkit; Lec. 4 turns it on the optimal-growth Bellman equation.

Summary

1. **One stack, layered:** AI evolved from rule-based systems through ML and deep learning to today's generative and agentic era.
2. **Three pillars of ML:** supervised (prediction), unsupervised (structure), reinforcement (sequential decisions) — with deep learning as the shared engine.
3. **Predictive AI fails** when the future is not in the data: COMPAS, Fragile Families — proxies, gaming, and drift; the Lucas critique in ML clothing.
4. **The generative stack:** NLP $\rightarrow$ LLMs $\rightarrow$ agents — systems that *generate* and *act*, already woven into everyday and research use.
5. **Generative AI fails differently:** it *can*, but unreliably or harmfully — hallucination, deepfakes; so always ask *what task, what data, what claim*.
6. **Next:** Lec. 3 goes hands-on with the dynamic-macro toolkit — why ML for economists, neural networks, training, and specialized architectures.