AI for Economic Research: Dynamic Models, Language, and Agents
Lecture 1.2: AI as a Research Collaborator

Zhigang Feng

2026-07-07 10:16:51

Lecture 1 in One Map

One lecture, three decks, one claim: AI is new *cognitive capital* — cheap, fast, non-rival — and this course trains you to deploy it on real research.

Act 1.2 (here) *shows the claim in action: DeepRamsey co-develops a dynamic model, GRAM shows context — not capability — is the bottleneck, the AI-for-research map places the tools (with causation as the standing guardrail), and a live measurement project asks what AI can actually finish.*

Case A — DeepRamsey

AI as co-developer of a dynamic model — and where human judgment stayed load-bearing

AI Compresses the Theory→Code Loop

Case study (full version: Lec. 10 T8): AI-assisted extension of the Feng, Han & Zhu (2026) deep Ramsey solver from a representative-agent to a *heterogeneous-agent* model — a new paper, Chien, Feng & Wen (in progress).

- The theory–code–debug loop compressed **from months to days**: drafting, refactoring, auditing, and diagnostics stop being bottlenecks.
- Documentation became *executable*: the TeX spec, audit notes, and code evolve in lockstep.
- **Where judgment stayed human:** the AI could not decide what was economically *interesting*. It *could* sometimes flag a mistimed equation or a non-credible number — but confirming that stayed the economist's job.

The AI supplies **throughput**; the economist supplies **judgment** — the central lesson of the case.

Based on Feng, Han & Zhu (2026), *Ramsey Taxation, Public Debt Sustainability, and Asset Pricing in Incomplete Markets*; HA extension: Chien, Feng & Wen (in progress).

Start From a Validated Seed — Audit Before Extending

The conversation did not start with "build me an HA model." It started with comprehension and audit tasks on the seed (the RA deep Ramsey solver):

1. *Restate* the model and the algorithm; list every equation and where it is implemented.
2. *Audit* the code against the TeX: report every mismatch (scoring, bounds, sampling, penalties); classify each as a **correctness risk** or a **documentation gap**.
3. *Refine* to a consistent document+code pair — add no new economics yet.

Why audit first? Extending an inconsistent seed propagates the inconsistency into every later version. The audit is the cheapest insurance in the whole project.

Extend on Paper, Prototype with AI, Let Failures Teach

Milestone ladder: refined RA → stripped-down HA model (*TeX first*) → AI-built code (`v1`) → full model (`v2`), where *debugging becomes architecture*.

Contribution	AI	Economist
Drafting/reorganizing LaTeX; vectorized PyTorch	✓	
Proposing penalty & boundary-learning patterns	✓	
Summarizing diagnostics & code/document mismatches	✓	
Checking economic correctness of formulations		✓
Deciding which simplifications are acceptable		✓
Catching conceptual bugs (timing, transition direction)		✓
Setting validation benchmarks; accepting results		✓

Three Discoveries the Failures Forced

Hitting the *same* failure, then asking *why*, imported three tools economists rarely reach for:

1. α-**shapes** (Edelsbrunner–Mücke 1994) to represent the *generically non-convex* feasible set from sampled points — surfaced when the boundary learner kept misclassifying near-kink states.
2. **Viability kernel** (Aubin 1991) *precedes* admissibility: the same monotone-set-operator structure as APS self-generation (Abreu–Pearce–Stacchetti 1990; Phelan–Stacchetti 2001; Feng 2015).
3. A **Lyapunov-type penalty** for saddle-path selection under the $\beta R = 1$ unit root — flat along the unit-root direction, punishing only the movement that signals divergence.

Pattern: an ad hoc fix fails → AI-generated diagnostics isolate *why* → the economist recognizes the deeper structure → joint design of the right tool. Cross-literature imports got *cheap*.

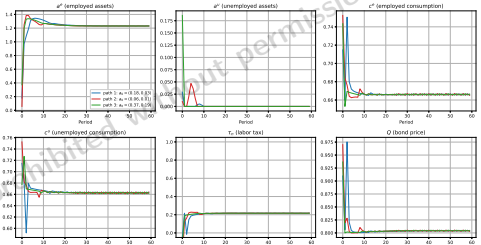
Result: A Bistable Long Run

Heterogeneity creates *two* long-run outcomes: full self-insurance (FSI), or **immiseration** (assets pile up, consumption collapses).

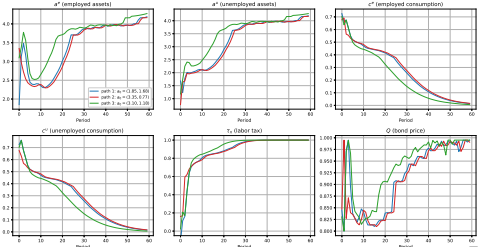
- Which one is reached depends on **where the economy starts**.
- CRRA: only $\approx 13\%$ of initial conditions (a thin low-asset sliver) converge to FSI.
- Basin boundaries are thin and perturbation-sensitive — mapped with the α-shape machinery.

Long-run outcomes are history-dependent: policy alone does not decide between self-insurance and immiseration — initial conditions do.

Full case study: Lec. 10 T8.



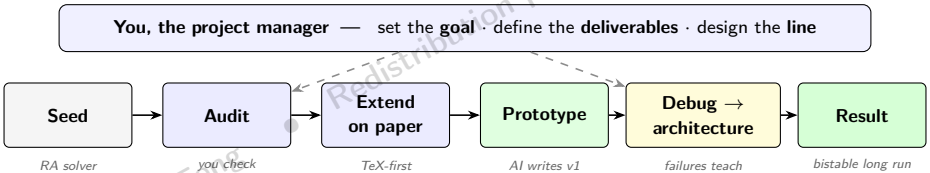
A path converging to full self-insurance.



The Researcher Becomes a Project Manager

As AI supplies the throughput, your job shifts from *doing every step by hand* (the old artisan mode) to *directing the line*. Like a plant manager, you own three things:

- **Vision** — a sharp research *goal*: what is the end product, and what makes it *interesting*?
- **Deliverables** — the intermediate steps and products the goal decomposes into (the milestones).
- **Workflow** — the *assembly line* from inputs to result: order the stations, and assign each to AI (throughput) or to yourself (judgment).



The Case A ladder you just saw *was* this line: audit → extend on paper → prototype → debug-into-architecture → validate. You never worked the conveyor — you **designed** it and stood at the **judgment gates**.

Case B — GRAM

Capability is not the bottleneck; *context* is

Capability Is Not the Bottleneck — Context Is

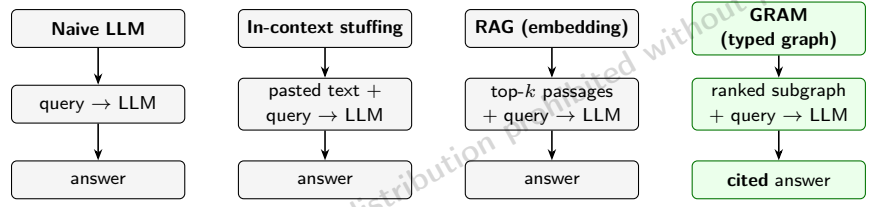
The same frontier tools, available to everyone, produce dramatically different research outcomes — because of how the human supplies *context*.

- Modern models are already capable enough for sophisticated quantitative-macro work...
- ...but **ungrounded**, they fail in characteristic ways:
  - confident *hallucinated* citations with real-looking DOIs;
  - plausible-but-wrong methods (valid elsewhere, invalid *here*);
  - knowledge-cutoff blindness (the actual algorithm is post-cutoff);
  - fluent, unsourced prose you cannot check.

The bottleneck is not raw capability — it is getting the *right field-specific context*, **with its conditions of validity**, in front of the model.

Based on Feng, Miao, Xie, Zhong & Zhou (2026), *GRAM: Generative Reasoning and Augmented Modeling for Quantitative Macroeconomics* (instructor WIP).

Four Ways to Give an LLM Context



Only GRAM adds **condition-aware retrieval**: a method is retrieved *together with the conditions under which it holds*. Fact lookup and compositional retrieval are not enough.

Why Naive RAG Falls Short: Economics Is Math *and* Narrative

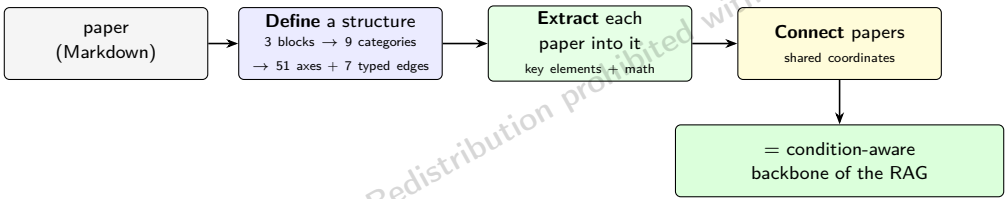
Two retrieval traditions each handle only *one* side of a paper. Economics papers are *both* math (the analytical scaffold) and prose (the model's anatomy and mechanisms) at once.

Math info retrieval	Social-science / legal RAG	Economics — GRAM
Pure-math papers	Qualitative prose (e.g. LegalGraphRAG)	*Both at once*
Unit: theorem / equation	Unit: case / statute / note	Unit: typed economic axes, *equations carried inside*
Structure from embedding similarity	Typed graph injects domain knowledge	Typed graph + textbook canon
Strong on math, blind to prose	*Strong on prose, no math*	*Math is analysis; prose is mechanism*

A flat embedding store indexes *surface text* and respects neither structure. GRAM **sorts out the structure**: one typed extraction carries both — the prose anatomy with its math inside — so retrieval matches the *conditions* a question needs, not just the most-similar passage.

PromptEcon: A Shared Basis for Models

The engine of Stage 3 — one fixed basis for the anatomy of *any* structural macro model:



The 51 axes are the unit of retrieval. Because every paper is projected onto the *same* coordinates, papers align by typed concept — and a query can demand the axis (e.g. `solution_method`) that fits its stated conditions.

Payoff — and Where This Goes

- **Replicate in days:** an empirical researcher opened an *unfamiliar* codebase, paired it with an agent, and in days replicated the model and solved computational problems the original author had been stuck on.
- $\sim$**10**$\times$ **throughput:** a computational researcher reached a roughly ten-fold improvement on *their own* code with an AI assistant in the loop.

Forward pointers: the retrieval mechanics (embeddings, RAG, GraphRAG) $\rightarrow$ Lec. 8; agentic orchestration of multi-step research $\rightarrow$ Lec. 9. Building such infrastructure is **Capstone Track D**.

AI-for-Research Map & the Causation Guardrail

Where AI helps across the research pipeline — and the one claim it must never make for you

AI Tools for Economic Research: A Map

Task	AI Approach	Economic Application
Prediction	Supervised learning	Forecasting, scoring, targeting
Pattern discovery	Unsupervised learning	Consumer types, factor models
Text analysis	NLP / LLMs	FOMC, speeches, filings
Causal inference	ML + IV / DiD	Heterogeneous treatment effects
Structural estimation	RL / DL	Solving complex macro models
Literature review	RAG + LLM	Systematic reviews, survey papers
Coding	Agentic AI	Data cleaning, estimation, tables

The rows above are the material covered in
苗建军、冯志钢：《面向经济学人的人工智能》，高等教育出版社，2027

AI in Economic Research: Concrete Examples (I)

- **Prediction (supervised).** Nowcast GDP / inflation from high-frequency and alternative data (card spending, satellite night-lights); predict firm default or earnings surprises from filings. (Lec. 3)
- **Pattern discovery (unsupervised).** Topic models / embeddings on FOMC minutes and earnings calls to surface latent themes; cluster firms or households into types from panel data. (Lec. 7)
- **Text as data (NLP / LLMs).** Score hawkish/dovish tone in central-bank language; extract firm-level risk exposure from 10-K Item 1A; build a policy-uncertainty index from newspapers (Baker–Bloom–Davis). (Lec. 7–8)

*The map's first rows, made concrete: AI turns messy inputs into **measures** and **datasets**.*

AI in Economic Research: Concrete Examples (II)

- **Causal inference (ML + IV/DiD).** Double/debiased ML to control high-dimensional confounders (returns to schooling); causal forests for heterogeneous treatment effects (who gains from a policy); synthetic control for a single treated unit (Abadie). (Lec. 3)
- **Structural estimation (DL / RL).** Neural-network solutions of high-dimensional dynamic models (Aiyagari, Krusell–Smith); deep learning for the Bellman / Euler equations where grids break down. (Lec. 4–6; **Capstone A**)
- **Literature & coding (RAG + agentic).** GraphRAG over NBER/arXiv for systematic surveys (the T1.2 GRAM case); an agent builds the data → clean → estimate → table pipeline and replicates a paper from public data. (Lec. 8–10; **Capstone C, D**)

The guardrail. AI supplies the *prediction, measurement, and computation*. Identification — *why* a number is causal — remains the economist's job: prediction $\neq$ causation.

AI Strengths and Limitations for Economists

Where AI Excels

- High-dimensional pattern recognition
- Text analysis at scale (FOMC, patents, filings)
- Constructing new datasets from unstructured data
- Automating repetitive coding and cleaning tasks
- Solving high-dimensional optimization problems (DL for macro models)

Where Economists Add Value

- Identification and causal reasoning
- Economic structure: what are the right state variables?
- Welfare analysis and normative judgment
- Designing the question AI answers
- Validating model outputs against economic theory

"AI compensates for human cognitive limitations in pattern matching but lacks the human capacity for causal

A Live Measurement Example

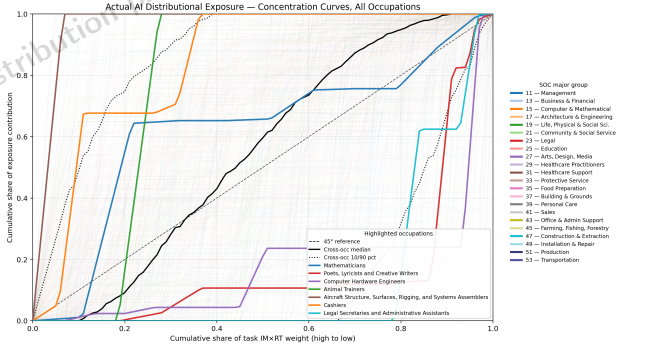
Instructor work-in-progress: measuring what AI can attempt vs. finish

Measuring AI as Cognitive Capital: From the Capability Index to Data

- **Bridge.** The task-level CES slide (T1.1) carried two unobservables: an AI-capability index θ_i and an adoption margin E_i. *Can we measure them?* Merge the **Anthropic Economic Index** (observed Claude usage) onto **ONET** tasks.
- **Per occupation:** sort tasks *most-important-first*, then trace cumulative AI-usage share against cumulative task-importance weight ⇒ a Lorenz-style deployment curve.

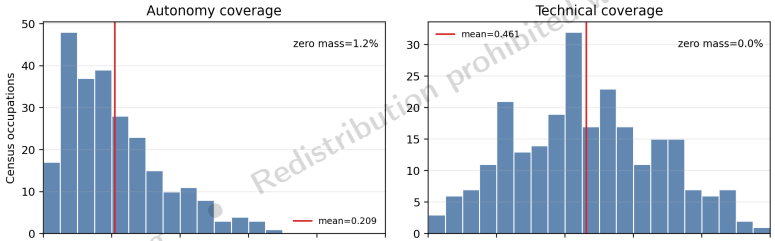
Three deployment indices:

- **Signed Gini** G: >0 = AI on core/important tasks; <0 = peripheral.
- $|G|$: *sharpness* of the human–AI task split (direction-free).
- $\bar{E}$: importance-weighted usage *intensity*.



Is AI Actually Capable Yet? A Task-Level Capability Index

- An LLM rates *every* ONET task on two axes (blinded, averaged over 3 passes):
  - $\mathrm{CAP}_{\mathrm{tech}}$ — technical coverage: can current AI *attempt* the task? (mean ~ 0.46)
  - $\mathrm{CAP}_{\mathrm{auto}}$ — autonomous coverage: can it *finish* without human review? (mean ~ 0.21)



AI can *attempt* much but *finish* unsupervised little — consistent with Acemoglu's "∼4.6% of tasks exposed."

Caveat: this is an LLM *self-rating* (blinded, 3-pass averaged) — a model judgment, not external ground truth.

Bridge: The Proof Is In — Now Make It Your Practice

Act 1.2 delivered the evidence. AI co-developed a dynamic model (DeepRamsey); context — not capability — emerged as the binding constraint (GRAM); and the task-level index showed AI can *attempt* much but *finish* unsupervised little.

- You have seen two live cases and one measurement project — with causation as the standing guardrail: AI drafts, computes, and searches; *you* own identification and the causal claim.
- **Next (1.3):** how this course works — the two-arc roadmap, the research-architect stance and tooling, and the replication-first capstone.

*Arc: 1.1 FRAME (cognitive capital) → **1.2 PROOF** (AI doing research) → 1.3 PRACTICE (course + capstone).*