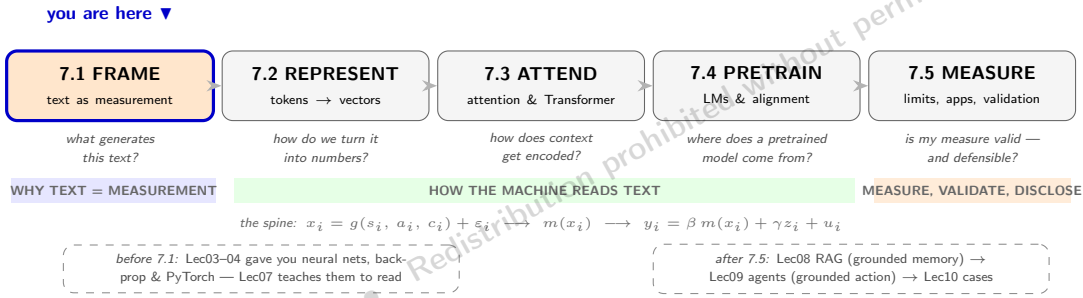
AI for Economic Research: Dynamic Models, Language, and Agents
Lecture 7.1: Text as Economic Measurement

Zhigang Feng

2026-07-30 11:20:36

Lecture 7 in One Map



Two case threads run through all five acts — **Thread A: the EPU index** *(Baker–Bloom–Davis 2016), the transparent dictionary baseline every method must beat;* **Thread B: FOMC hawk–dove** *(central-bank communication), the running example for embeddings, attention, and validation. Watch for them.*

Text as Data

Why language became an economic measurement problem

The Opening Hook: Reading Uncertainty out of Newspapers

In 2016, three economists asked: can you read policy uncertainty out of the newspaper?

- Baker, Bloom & Davis (2016, *QJE*): scan 10 major US newspapers, month by month, and count articles containing **all three** of {`economic`/`economy`}, **and** {`uncertain`/`uncertainty`}, **and** one of {`Congress`, `deficit`, `Federal Reserve`, `legislation`, `regulation`, `White House`}.
- Scale by each paper's total output, standardize, average across papers, index to 100:

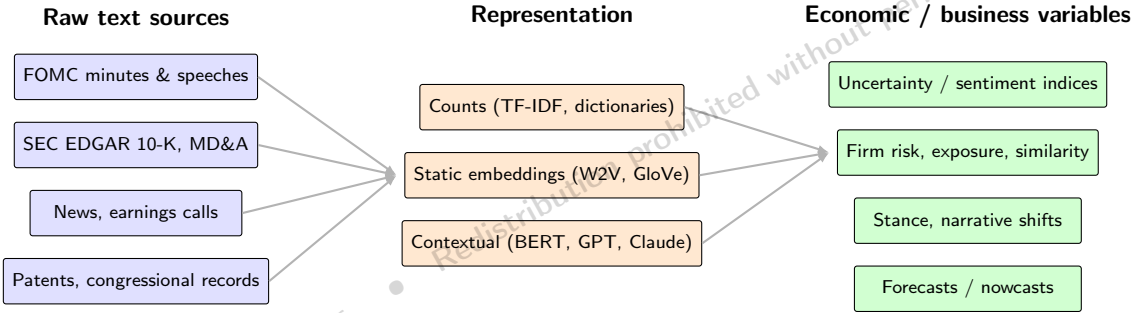
$$\mathrm{EPU}_\tau \;=\; 100 \times \frac{u_\tau}{\bar{u}}, \qquad u_\tau \;=\; \frac{1}{10}\sum_{p=1}^{10} \frac{c_{p\tau}/N_{p\tau}}{\sigma_p}.$$

where $p = 1, \ldots, 10$ indexes newspapers, τ the month; $c_{p\tau}$ = qualifying articles in paper p; $N_{p\tau}$ = paper p's total articles (dividing removes size differences); σ_p = std. dev. of paper p's scaled series (so no one paper dominates); u_τ = 10-paper average; $\bar{u}$ = its sample mean, so $\times 100$ sets the average to 100.

- No statistical agency publishes "policy uncertainty" — it is not directly observed anywhere. Yet in a VAR, EPU spikes precede investment and employment declines.

The move that matters: unstructured language became a *regression-ready variable* — this act asks what that takes, and where it goes wrong. *(Thread A begins.)*

Text as Economic Data — The Big Picture



The lecture builds the middle column rung by rung: Act 2 counts → static embeddings, Act 3 contextual machinery, Act 4 how those models are trained — and Act 5 turns the machinery into the right-hand column, with validation.

Why Economists Care About LLMs

- **Text as data** — modern economic research increasingly leverages textual sources:
  - Corporate filings (10-K, 10-Q, MD&A)
  - Central bank speeches, FOMC minutes, ECB statements
  - Earnings calls, analyst reports, regulatory filings
  - News, social media, congressional records

- **Applications:**
  - Sentiment analysis for financial forecasting
  - Policy stance detection and uncertainty indices
  - New text-derived macroeconomic indicators
  - Summarization of long, technical documents

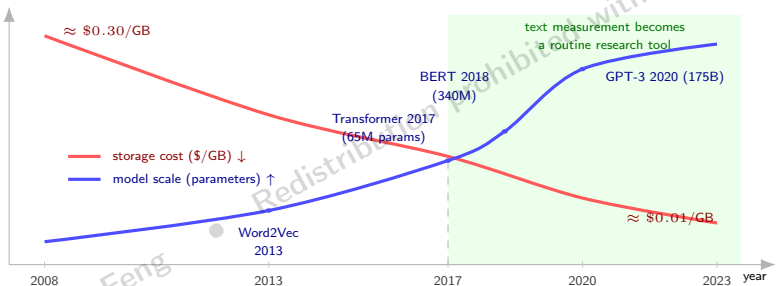
- **Why LLMs in particular:**
  - Capture nuanced language and domain-specific jargon
  - Handle long-range context (sentences, paragraphs, whole documents)
  - Provide a unified architecture for many tasks

Three Roles of Text in Economic Research

Role	What text is	Example
Measurement tool	A way to construct previously unobserved variables	Baker–Bloom–Davis (2016) Economic Policy Uncertainty index from newspapers
Behavioral outcome	A response variable revealing decisions or institutional change	Hansen, McMahon & Prat (2018): how FOMC transparency reform changed committee speech
Treatment variable	A quasi-experimental input affecting behavior	Central bank communication shocks and inflation expectations

Why Now? The Feasibility Window

Two curves crossed around 2017: the cost of handling text collapsed while model capability exploded.



Both quantities on log scales; curves are stylized interpolations through the marked real data points. Storage costs after McCallum's disk-price series; model sizes from Vaswani et al. (2017), Devlin et al. (2018), Brown et al. (2020).

The Spine

A data-generating process for text — and the two hazards it creates

An Identification-First View of Text

- **Engineering perspective first:** to a machine, a document is just a character string. But every string was *produced* — by a person, in a situation, for an audience. Before any econometrics, ask: what generated this string?

- **The data-generating process (DGP):**

$$x_i \; = \; g\big(s_i,\, a_i,\, c_i\big) \; + \; \varepsilon_i$$

  - s_i: the latent **economic state** we want to measure (policy stance, sentiment, firm risk)
  - a_i: the speaker's **strategic framing** — audience management, spin, emphasis, tone chosen for effect
  - c_i: **institutional context** — venue, legal disclosure requirements, format conventions

- **Why this matters for a downstream regression** $y_i = \beta\, m(x_i) + \gamma z_i + u_i$: if the strategic framing a_i is correlated with the regression error u_i, then the text feature $m(x_i)$ is *endogenous* — the familiar omitted-variable problem, now hidden inside the language.

Two Firms, Same Risk, Different Words

The endogeneity hazard, made concrete. Imagine two firms facing the *same* objective political risk s_i.

- Firm A's management discloses risks proactively — building a reputation for transparency.
- Firm B's management avoids politically sensitive topics — afraid of spooking the market.
- Same s_i, different strategic framing $a_i \Rightarrow$ different texts $x_i \Rightarrow$ different measured risk $m(x_i)$.

- If the disclosure choice a_i correlates with anything in u_i (governance quality, investor base, financial health), $m(x_i)$ is **endogenous** — before any fancy NLP even enters.

Setting from Hassan, Hollander, van Lent & Tahoun (2019, QJE).

EPU: The Validation Standard (Thread A)

Why is a word-count index taken seriously? Because every step of it was audited.

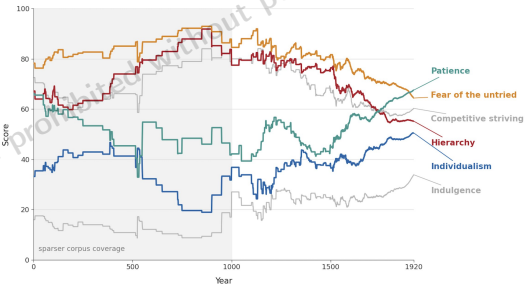
- **Human audit:** a trained team hand-coded 12,009 articles (1900–2012) under a 65-page guide; the six policy terms were selected from ∼32,000 candidate combinations to minimize false positives + false negatives.
- **Machine vs. human:** the "machine" is the automated word-count index itself — a program that flags every article matching the six-term rule and tallies the monthly counts, with no human judgment in the loop. Run over the same years the team hand-coded, it tracks the human-coded series at correlation 0.86 (quarterly, 1985–2012) and 0.93 (annual, 1900–2010); the gaps are uncorrelated with GDP growth.
- **Robustness:** dropping any one policy term barely moves the series; a 1,500-newspaper daily version correlates 0.85 with the 10-paper index.
- **And still — the identification critique:** does EPU *cause* downturns, or do downturns make journalists write "uncertainty"? The VAR ordering (EPU first) is a maintained assumption, not an identification argument.

Source: Baker, Bloom & Davis (2016, QJE).

From Counting to Reading: Culture out of 23,000 Books

EPU fixes a six-term dictionary in advance and counts. What if you ask the model what each book endorses instead?

- **Corpus:** 23,000+ books of the Western canon, read *in context* — not searched for keywords.
- **The judgment asked of the model:** for each of **six** cultural dimensions (adapted from Hofstede 2001), does this work **endorse**, **reject**, or **merely depict** that position? Passage-level score $x_d \in \{-1, 0, +1\}$.
- **From flow to stock:** scores accumulate like capital, $\delta_K = 0.005$ — a 99.5% annual retention rate, half-life ≈ **138 years**. Culture is inherited, not consumed.
- **One summary number:** the *Innovation Wedge* $= (\mathrm{PD} + \mathrm{UA})/(\mathrm{INDIV} + \mathrm{LTO})$ — deference and fear of the untried, over autonomy and a long horizon. It **falls 51%** between 1000 and 1920.



Six dimensions, year 0–1920. The shaded pre-1000 band is sparser corpus coverage — shown for direction only, not for inference.

Source: Stefanski (2026), "Cultural Capital and the Productivity of Ideas: Evidence from Historical Texts," Univ. of St Andrews Economics Discussion Paper No. 2602. Figure reproduced with attribution.

Why It Passes the T1 Bar

The EPU frame set the standard: audit every step. This case meets it with a different machine.

- **Validation.** Experts *blinded* to the scores agree with the score-implied direction in **93.0%** of anchor comparisons among books they know; the stocks reproduce cross-country orderings from modern surveys; and an **independent 5,000-book archive** on a frozen pipeline reproduces the decline.
- **The DGP, instantiated.** In $x_i = g(s_i, a_i, c_i) + \varepsilon_i$: s_i is the work's stance; a_i the authorial framing — a satire *depicts* what it does not endorse, which is why the scale needs a "depicts" category and a dictionary cannot supply one; c_i is the archive.
- **What it buys.** The falling wedge raises 1920 productivity to **1.78**$\times$ its counterfactual level, explains **two-thirds** of the first sustained growth acceleration (1500–1700), and accounts for **38.7%** of 1920 productivity growth — culture as a *measured series*, not a residual.

c_i **binds here.** The corpus is **"selected twice"**: history chose what was preserved and taught; digitizers chose which survivors were scanned. The first selection is part of the object of study — the second is not, and no amount of model quality repairs it.

Cost drove the architecture: Gemini 2.5 Pro runs the context pass once per book, Flash the passage stages that run hundreds of thousands of times.

China Parallel: The Same Method on Chinese History

The method is not about the West — it is about any canon large enough to read and dated well enough to index.

- **Corpora already digitized:** the Twenty-Four Histories; the *Siku Quanshu*; local gazetteers (*difangzhi*) — thousands of volumes at fine geographic resolution; Ming–Qing memorials and legal archives; examination essays; clan genealogies; *Shenbao* (1872–1949) for the late period.
- **The question it could reach.** The Needham puzzle and the Great Divergence turn on cultural claims — attitudes to hierarchy, novelty, the long horizon — argued narratively because no series exists. The same scoring would put a Chinese series beside the Western one and let the growth model arbitrate instead of the anecdote.

What breaks — and this is the point of the exercise.

- **Selection is stronger here, not weaker.** Imperial editorial projects did not merely preserve — the *Siku Quanshu* suppressed and altered as it collected. c_i is partly a *censorship* process, correlated with the very attitudes to authority being measured.
- **Register drift.** Classical and vernacular coexist for centuries; the same characters shift meaning across them.
- **Segmentation is itself a modeling decision**, before any scoring (Act 2).
- **Hofstede's dimensions were built for modern survey respondents.** Porting them to an imperial canon is an assumption needing its own blinded anchor study — the 93% exercise, repeated with Sinologists.

Preview: The Machine Ahead

One equation will power everything in Acts 2–4

A Brief History: RNN → Transformer → GPT

- **2003:** Bengio et al. neural language model — continuous word vectors
- **2013:** Mikolov et al. *Word2Vec* — dense word embeddings at scale
- **2014:** Pennington et al. *GloVe* — global co-occurrence matrices
- **2017:** Vaswani et al. *"Attention Is All You Need"* — the Transformer replaces RNNs/LSTMs
- **2018:** Devlin et al. *BERT* (encoder); Radford et al. *GPT* (decoder)
- **2020–2022:** GPT-3, instruction tuning, RLHF → ChatGPT moment
- **2023–2026:** GPT-4/5, Claude (3–4), Gemini, LLaMA-3, DeepSeek; long-context (100k–1M tokens), open-weights ecosystem, agentic systems

Driver of the shift: attention scales better than recurrence, both in compute (parallel) and in quality (long-range dependencies).

The Autoregressive Principle

- An LLM is fundamentally a **next-token predictor**:

$$p\big(t^{(n+1)} \mid t^{(1)}, t^{(2)}, \ldots, t^{(n)}\big).$$

- **Generation:** sample $t^{(n+1)}$ from this distribution, append to the context, and repeat.
- **What this single equation enables** (the full revisit comes at the end of Act 4, once pretraining and fine-tuning are on the table):
  - **Summarizing 10-K filings or FOMC minutes** — prompt with document + instruction; the model predicts summary tokens one by one
  - **Generating narrative forecasts** conditional on macro state — condition on data in the prompt; next tokens extend the narrative
  - **Producing scenario analyses in plain language** — the scenario is the context; prediction fills in the implications
  - **Writing and debugging code** for empirical work — same objective over a code corpus yields code completions
- **Key insight:** "everything is next-token prediction" is mathematically tiny but pedagogically and practically enormous.