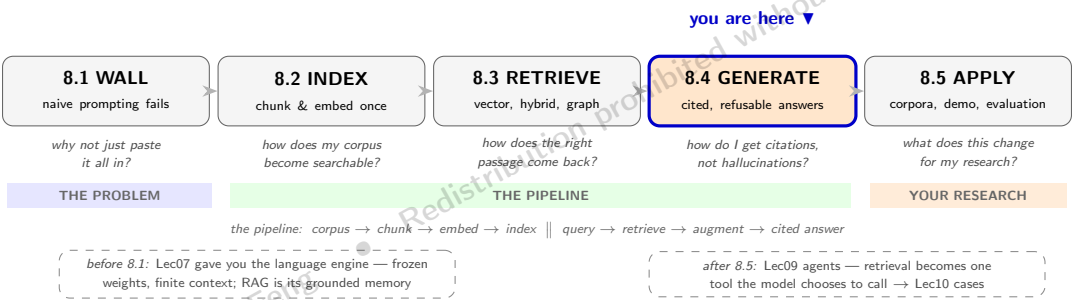
AI for Economic Research: Dynamic Models, Language, and Agents
Lecture 8.4: Generate — Grounded Answers and RAG vs Fine-Tuning

Zhigang Feng

2026-07-10 19:45:01

Map: Act 8.4 — From Evidence to a Cited Answer

Retrieval hands back passages; this act turns them into an answer you can put in a paper.



Route this deck: the generation contract (ground, cite, refuse) → a worked FOMC run → the decision matrix vs fine-tuning. Thread A carries the worked example; Thread B returns in the Act 8.5 classroom demo.

Grounded Generation

From retrieved evidence to a cited, refusable answer

From Retrieved Chunks to a Cited Answer

The generation contract: the answer must be grounded in the retrieved evidence, cited chunk by chunk, and refusable when the evidence runs out.

- We now have the top-K chunks for the query (typically $K = 3$ to 10). **Insert them into a prompt template and send to the LLM.** The model sees:
  1. System instructions (how to behave)
  2. The retrieved chunks (the evidence)
  3. The user question
  4. Output-format instructions (how to cite)

- This is **the only place a generative model is used** in the whole pipeline. Everything before — chunking, embedding, retrieval — exists to put high-quality evidence in front of the model.

- **The single most important design choice** at this stage: the *prompt template*.

Prompt Template — Anatomy

```
SYSTEM:
You are a research assistant for academic economists.
Answer the user's question using ONLY the context below.
If the context does not contain the answer, reply
exactly: "I don't know based on the provided sources."
Cite sources inline using [doc_id:page] notation.

CONTEXT:
[1] (fomc_2024_06:p3)
"Participants observed that inflation remained elevated..."

[2] (fomc_2024_06:p4)
"Several noted that core services prices, especially shelter,
continued to rise..."

QUESTION:
How did the FOMC discuss inflation expectations in June 2024?

INSTRUCTIONS:
Answer in 2-3 sentences. Cite each claim inline as [doc_id:page].
```


Four parts: system contract, evidence, question, format — the contract lines do the grounding work.

Source Attribution Mechanics

Every claim in an academic paper must be traceable to a source — an ungrounded LLM output is not citable; a grounded output is.

- **How to enforce:**
  1. At chunking time, give every chunk a stable `doc_id` and `page` (or `turn`, or `section`).
  2. At retrieval time, return the metadata along with the chunk text.
  3. In the prompt template, label each chunk `[1] (doc_id:page)` and instruct the model to cite using that label.
  4. In post-processing, replace inline citations with hyperlinks back to the original document.

- **Before / after:**
  - *Ungrounded:* "The FOMC was concerned about inflation persistence in 2024."
  - *Grounded:* "The FOMC was concerned about inflation persistence in mid-2024 `[fomc_2024_06:p3]`, particularly in core services prices including shelter `[fomc_2024_06:p4]`."

Refusal and Faithfulness

"I don't know" is a feature, not a bug: in research, a fabricated answer is far worse than a refusal.

- **Faithfulness** (every claim supported by the retrieved context) matters more than *fluency*. A fluent fabrication is the worst possible outcome for research.
- **How to enforce refusal:**
  - Explicit system-prompt instruction: "If the context does not contain the answer, reply exactly: I don't know based on the provided sources."
  - Low generation `temperature` (0.0–0.2) to discourage creative inference.
  - In post-processing, run a faithfulness check (Act 8.5) and flag unsupported claims.
- **Example refusal (good behavior):**
  "I don't know based on the provided sources. The retrieved minutes from June 2024 do not discuss the impact of fiscal policy."

This is what protects the researcher from quoting fiction.

Worked Example: Full Pipeline Run

Query. *"How did Powell discuss inflation expectations in 2023 press conferences?"*

Pipeline trace:

1. **Embed query** with `bge-large-en-v1.5` → 1024-d vector.
2. **Retrieve top-50** via Chroma + BM25 hybrid. Filter: `doc_type = press_conference` AND `year = 2023`.
3. **Re-rank** the 50 with `bge-reranker-large` → top-5.
4. **Inject top-5 chunks** into the prompt template (previous slide), tagged `[powell_2023_03:p4]`, `[powell_2023_06:p2]`, ...
5. **Send to the generator LLM** at `temperature=0.1`. Receives:
   > *"In 2023 press conferences, Chair Powell repeatedly emphasized that long-run inflation expectations remained well anchored [powell_2023_03:p4], while noting that short-run expectations had risen with realized inflation [powell_2023_06:p2]. By year-end, he highlighted that survey- and market-based measures had begun to ease [powell_2023_12:p3]."*
6. **Post-process:** convert inline citations to clickable links back to the transcript pages.

RAG vs. Fine-Tuning

Facts need retrieval; style needs weights

Decision Matrix: RAG vs. Fine-Tuning

Dimension	RAG	Fine-tuning
Knowledge freshness (update daily)	strong	weak
Source attribution / citations	strong	none
Reduces hallucination	strong	limited
Cost to add new documents	near zero	re-train
Latency at query time	moderate	low
Style adaptation ("write like the Fed")	weak	strong
Output-format adaptation (structured JSON)	limited	strong

Reading the matrix (green = native strength, yellow = workable, red = fundamental weakness): RAG dominates on the four dimensions that matter most for *research* — freshness, attribution, hallucination control, update cost. Fine-tuning wins on style and format, which matter for *production* but rarely for research.

When Fine-Tuning Still Makes Sense

Fine-tuning earns its keep on style, format, and jargon — not on facts.

- **Style transfer.** You want the model to *write like a Federal Reserve economist* — cadence, hedging, phrasing. Fine-tune on a thousand FOMC-minutes examples. RAG can't teach style.
- **Structured-output formatting.** Always emit a JSON dictionary of DSGE parameters? Fine-tune on ∼500 input/output pairs.
- **Specialized vocabulary.** Teach the model that `LSAP` = Large-Scale Asset Purchase, `NIPA` = National Income and Product Accounts.
- **Latency-critical settings.** No retrieval round-trip — useful in real time, rarely binding for research.

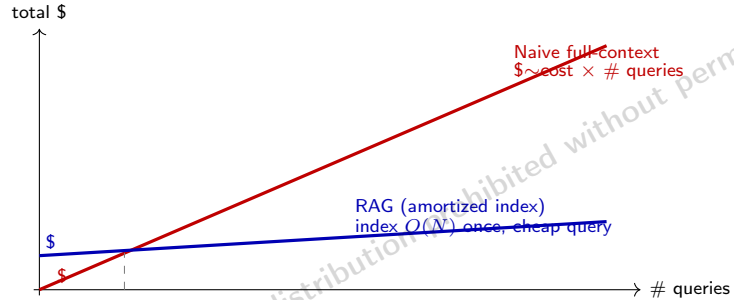
Crucial caveat: fine-tuning ≠ teaching new *facts*. Models memorize facts unreliably during fine-tuning and still hallucinate. *If you want the model to know new facts, use RAG* — Lec. 7 T5b's closing rule: fine-tuning stores *style*, not *facts*.

The Hybrid Pattern

In production the modern answer is both, in different roles: fine-tune the small model, retrieve for the big one.

- **Fine-tune the embedding model** on a domain corpus (say, 100K FOMC sentences with synthetic query–document pairs). Cheap, and it improves retrieval quality dramatically.
- **Use a general-purpose frontier LLM as the generator** — no fine-tuning of the big model. **RAG is the glue:** the fine-tuned embedder finds the right chunks; the general LLM writes fluent, faithful prose from them.
- **Why this works:** domain understanding from the (small) fine-tune, freshness + attribution from RAG — without the cost or fragility of fine-tuning a frontier-scale generator.
- The dominant production architecture (mid-2026) for LLM applications over specialized corpora — legal, medical, financial, academic.

Cost: Index Once, Query Many (Visual)



- **Naive full-context prompting:** cost scales *linearly* with context size *and* # queries — every question repays the full corpus token bill (Act 8.1's meter wall).
- **RAG:** pay $O(N)$ embedding calls *once* at index time; each query ships only the top-K chunks ($\sim$1–5% of the corpus). Marginal cost per query collapses.
- **Per-token rates and the reproducibility tax:** Lec. 7 T4, "Tokens, Cost, and Reproducibility." Past a handful of queries, index-once dominates on total spend *and* latency.

Bridge: The Pipeline Is Complete — Now Make It Earn Its Keep

Act 4 closed the loop: retrieved chunks become cited evidence, the generator is contractually forced to ground, cite, and refuse — and the decision matrix says when to fine-tune instead.

- **What changed:** the output is now *auditable*. That is what makes RAG a research tool rather than a chat trick.
- **Rule to remember:** style, format, jargon → fine-tune; facts → retrieve.
- **Next — Act 8.5 (APPLY):** the payoff. Six research corpora end to end, a classroom OLG demo you can run offline today, the six failure modes and their mitigations, RAGAS + human validation — and the bridge to Lecture 9's agents.

Arc: Act 1 WALL → Act 2 INDEX → Act 3 RETRIEVE → **Act 4 GENERATE** *→ Act 5 APPLY.*