

RAG for LLMs

Daniil Sherstnev

February 2026

- ▶ What is RAG?
- ▶ Why RAG?
- ▶ How does RAG work?
- ▶ Code walkthrough
- ▶ Examples

- ▶ In-context learning (ICL): provide examples to guide the model (3-6)
- ▶ RAG: providing dynamic, up-to-date, and proprietary data in the prompt for high-accuracy, verifiable responses
- ▶ Fine-tuning: updates model weights to learn domain-specific

RAG leverages external knowledge sources, advanced search techniques, and sophisticated language models to bridge the gap between static language models and dynamic, up-to-date information.

Sometimes RAG is called

- ▶ external knowledge base you add on top of LLM to search info in a customizable way
- ▶ nonlinear memory layer on top of LLMs

2 Why RAG? Do people use it?

| 3

- ▶ NYU Langone Health accuracy [jumped](#) past 58% and continues climbing vs regular 40% (Aug, 2025)
- ▶ Indegene's RAG GenAI solution [drives](#) 60% faster reporting for a global pharma leader
- ▶ DoorDash [enhanced](#) delivery support chat-bot using RAG (Sep, 2024)
- ▶ Bell and [Arcane](#) enhance its knowledge management processes and ensure its employees have access to up-to-date company and international policies

Companies use it because it is [cheap](#) (no need to train), it gives an option to feed the model with [up-to-date information](#), you can [deploy it locally](#)

3 How does RAG work? Math intro

| 4

Let Z be retrieved document with chunks z_i and x being a query, namely, the question we ask the model.

- 1 Choose pretrained embedding model and build $e_i = f_{emb}(z_i)$,
 $\mathbf{q} = f_{emb}(x)$
- 2 Use FIASS to set the distance rule $d(\mathbf{a}, \mathbf{b}) = \|\mathbf{a} - \mathbf{b}\|_2^2$
- 3 Select top-k highest similarity chunks as indices = $\operatorname{argmin}_i \|\mathbf{q} - \mathbf{e}_i\|_2^2$

In general,

$$P(z|x) = \frac{\exp(sim(x, z))}{\sum_j \exp(sim(x, z_j))}$$

3 Is it a nonlinear-memory layer?

| 5

Regular LLM:

$$P(y|x) = \prod_{t=1}^T P(y_t|y_{<t}, x)$$

Once we build $P(z|x)$, the model becomes:

$$P(y|x) = \sum_{z \in \mathcal{Z}(x)} P(z|x) \prod_{t=1}^T P(y_t|y_{<t}, x)$$

So the retrieved documents influence each token's generation through conditioning. Highly depends on retriever $P(z|x)$. If your question x is irrelevant to your base $Z \Rightarrow$ hallucinations or junk and no use

3 Parameters to care about

| 6

- ▶ **Chunk size** = how much text is stored in one retrievable unit z_i
 - > Small chunks: high precision, fine-grained retrieval, retrieve exact paragraphs or claims
 - > Chunk is too small \Rightarrow meaning may be fragmented
 - > Large chunks: high recall, coarse retrieval, retrieve broader context
 - > Chunk is too large \Rightarrow irrelevant text comes along
 - > **Recommended size**: 150–250 words for Methods / identification, 250–400 words for Abstracts / intro, 100–200 words for Tables / results
- ▶ **k** = how many chunks are retrieved and passed to the LLM
 - > Small k \Rightarrow focused, precise answers
 - > Large k \Rightarrow broader synthesis
 - > Noise vs signal trade-off
 - > **Recommended k**: 2–3 for Fact lookups, 3–4 for Citation grounding, 4–6 for Academic synthesis

7 papers published in Labor Economics in 2025. Uploaded to RAG publicly available working versions from 2022-2024

- ▶ David J. Deming: "Why Do Wages Grow Faster for Educated Workers?"
- ▶ Daniel Kopp: "Do Recruiters Penalize Men Who Prefer Low Hours? Evidence from Online Labor Market Data"
- ▶ Altoni, Giraldo Páez, Hynsjö, Vidangos: "Earnings, Marriage, and the Variance of Family Income by Age, Gender, and Cohort"
- ▶ Haddad & Kattan: "Female-Specific Labor Regulation and Employment: Historical Evidence from the United States"
- ▶ Kuhn & Osaki: "When is Discrimination Unfair?"
- ▶ Caetano, Caetano, Nielsen, Sanfelice: "The Effect of Maternal Labor Supply on Children: Evidence from Bunching"
- ▶ Borges, Estevan, Morin "Exam Reward Structure, Gender Performance Gaps, and Labor Market Outcomes"

4 Code walkthrough. Embedding

| 8

Function choice

```
def embed(text: str) -> np.ndarray:
    response = openai.embeddings.create(
        model="text-embedding-ada-002",
        input=text
    )
    return np.array(response.data[0].embedding, dtype="float32")
```

Run embedding and set FAISS

```
for c in chunks:
    text = c["text"].strip()
    if not text:
        continue
    emb = embed(text)
    embeddings.append(emb)
    valid_chunks.append(c)

emb_matrix = np.vstack(embeddings)
dimension = emb_matrix.shape[1]
index = faiss.IndexFlatL2(dimension)
index.add(emb_matrix)
chunks = valid_chunks
```

4 Code walkthrough. Similarity calculation

| 9

Here we choose k :

```
def retrieve(query: str, k: int = 5):
    q_emb = embed(query).reshape(1, -1)
    distances, indices = index.search(q_emb, k)
    return [chunks[i] for i in indices[0]]
```

And after that set the system role, model, temperature:

system: "You are an Economics Department Professor that gives clear, structured academic answers."

5 Example 1: GPT-5.1

| 10

question = "Which commonly studied labor-market topics are not addressed in the uploaded papers? Justify your answer using the documents."

However, several **commonly studied labor-market topics are not addressed** in any of these works:

1. **Unemployment dynamics and labor-market frictions** – none of the papers study unemployment duration, job-finding rates, or search and matching processes that are central in labor economics. The closest is Deming (2023), which looks at job mobility but not unemployment itself  doc_1.
2. **Wage-setting institutions and minimum wages** – there is no analysis of unions, collective bargaining, or minimum-wage policy, which are major topics in empirical labor economics.
3. **Labor demand and firm-side behavior** – apart from Deming's occupational sorting and Kopp's recruiter preferences, the firm production side, training decisions, or technology adoption's labor demand effects are not treated in depth.
4. **Immigration and labor-market integration** – none of the papers examine migrant workers, a staple area of modern labor research.
5. **Inequality and redistribution policy evaluation** – although Altonji et al. (2024) decompose income variance, none study taxation, social insurance, or labor-supply responses to transfers or unemployment insurance.
6. **Education policy and returns to schooling interventions** – while several papers relate to education indirectly (e.g., Deming 2023, Borges et al. 2024), none analyze schooling reforms or education policy as a causal determinant of labor-market outcomes.

In short, the uploaded papers emphasize **human capital formation, gender differences, discrimination, and family interactions**, but omit classic topics like **unemployment, labor institutions, migration, and policy evaluation**, which are foundational in the broader labor-economics literature.

5 Example 2: GPT-4o

| 11

question = According to the uploaded documents, what new empirical strategies have been introduced to study labor supply? Do not mention strategies not explicitly described in the papers.

The documents introduce several **new empirical strategies** for studying labor supply:

1. Control Function Approach with Bunching

One paper proposes a novel strategy using *bunching at zero hours* worked to identify the causal effect of maternal labor supply on child outcomes. By exploiting the fact that many mothers choose not to work (i.e., bunch at zero hours), the authors use the variation in children's cognitive skills among these non-working mothers to infer unobserved confounders. This method allows correction for endogeneity **without relying on instrumental variables**, expanding the usable sample to include more heterogeneous families and enabling finer-grained analysis of labor supply effects [doc_6](#).

2. Difference-in-Differences with Staggered Adoption

Another study uses a **staggered difference-in-differences** (DiD) design to evaluate the labor supply effects of historical female-specific labor regulations in the U.S., such as seating, health and safety, and night-work laws. By leveraging variation across states and time in the adoption of these laws, the authors estimate causal effects on female gainful employment while addressing potential confounders and testing for dynamic policy impacts [doc_4](#).

3. Dynamic Decomposition Using Structural Statistical Models

A third empirical strategy is seen in the use of a **structural statistical model** that integrates earnings, labor force participation, marriage, marital sorting, and fertility. The model decomposes the variance in family income by age, gender, and cohort using PSID data, allowing researchers to isolate the contribution of labor supply and employment shocks over the lifecycle [doc_3](#).

These approaches provide new methodological insights for identifying and estimating labor supply responses and their broader economic implications. 

5 Example 2: GPT-4o RAG

| 12

1. ****Instrumental Variables (IV) Strategy**:** - The documents describe an IV strategy that leverages exogenous variation in the diffusion of female-specific labor regulations across states within the same region. This approach uses regional waves of female-specific labor law passage as instruments for state-level law passage, following the methodology outlined by Acemoglu et al. (2019). The instrument is constructed based on the earliest year of passage of a female-specific labor law within each specific region (Appendix Table A8).

2. ****Control Function Approach**:** - A novel control function approach is introduced, which does not require the use of instrumental variables. This method leverages the fact that maternal labor supply is bunched at zero, addressing the endogeneity of maternal labor supply without relying on traditional IVs (Caetano, Caetano, and Nielsen).

These strategies aim to address endogeneity issues in labor supply studies, particularly concerning maternal labor supply and its impact on childhood skill development.

****References**:** - Acemoglu et al. (2019) - Caetano, C., G. Caetano, and E. Nielsen (2021, 2022)

