RAG Architecture

Theme: Pipeline stages, representation, and control

1. Pipeline Separation & Flow Control

(When things happen)

- **Indexing phase vs query phase**

  Conceptual role: lifecycle separation, offline vs online concerns

---

2. Representation & Granularity

(What units of knowledge look like)

- **Chunking strategies (size, overlap, hierarchy)**
- **Embedding selection tradeoffs**

  Conceptual role: how knowledge is *encoded* and *addressed*

---

3. Retrieval Quality Control

(What gets selected and refined)

- **Retriever vs reranker**

  Conceptual role: coarse recall vs fine precision

---

4. Context Assembly & Constraints

(What the model actually sees)

- **Context window management**

  Conceptual role: budget allocation under token constraints

Advanced RAG Topics

Theme: Reasoning depth, signal fusion, and query adaptation

1. Retrieval Signal Fusion

(How signals are combined)

- **Hybrid retrieval (BM25 + embeddings)**

  Conceptual role: symbolic + semantic alignment

---

2. Reasoning Depth & Dependency Handling

(How many steps of retrieval are needed)

- **Multi-hop retrieval**

  Conceptual role: compositional reasoning over documents

---

3. Query Adaptation & Reformulation

(How queries are improved)

- **Query rewriting**

Conceptual role: aligning user intent with retrievable representations

---

4. Context Validation & Grounding

(Ensuring relevance and traceability)

- **Context filtering and citation grounding**

Conceptual role: precision enforcement and trust

---

5. Data Modality Integration

(What kinds of sources are accessed)

- **Structured + unstructured retrieval**

Conceptual role: schema-aware + free-text fusion

Failure Modes

Theme: Retrieval breakdowns and knowledge misalignment

1. Generation–Retrieval Misalignment

(Retrieval worked, generation failed)

- **Hallucination despite retrieval**

Root issue: model ignores or misuses context

---

2. Coverage Failures

(Right answer exists, but not surfaced)

- **Missing context**
- **Over-retrieval**

Root issue: recall vs precision imbalance

---

3. Knowledge Freshness Failures

(Answer is outdated)

- **Stale knowledge**

  Root issue: index update or re-embedding gaps

---

4. Source Consistency Failures

(Multiple truths collide)

- **Conflicting sources**

  Root issue: lack of reconciliation or ranking logic

Why this structure works (important)

- Each subgroup answers **one dominant question**:
  - *When?* → pipeline
  - *What form?* → representation
  - *How selected?* → retrieval control
  - *Why wrong?* → failure class
- No topic appears in more than one conceptual bucket
- This maps cleanly to:
  - **system design discussions**
  - **debugging conversations**
  - **interview explanations**
  - **course/module layouts**

If you want, next we can:

- compress this into a **one-page mental map**
- align it to **RAG evaluation metrics**
- or map each subgroup → **design knobs + tradeoffs**

Just say the word.