Question-Answering (RAG)

One of the most common use-cases for LLMs is to answer questions over a set of data. This data is oftentimes in the form of unstructured documents (e.g. PDFs, HTML), but can also be semi-structured or structured.

The predominant framework for enabling QA with LLMs is Retrieval Augmented Generation (RAG). LlamaIndex offers simple-to-advanced RAG techniques to tackle simple-to-advanced questions over different volumes and types of data. You can choose to use either our prebuilt RAG abstractions (e.g. query engines) or build custom RAG workflows(example guide).

RAG over Unstructured Documents

LlamaIndex can pull in unstructured text, PDFs, Notion and Slack documents and more and index the data within them.

The simplest queries involve either semantic search or summarization.

- **Semantic search**: A query about specific information in a document that matches the query terms and/or semantic intent. This is typically executed with simple vector retrieval (top-k). Example of semantic search

- **Summarization**: condensing a large amount of data into a short summary relevant to your current question. Example of summarization

QA over Structured Data

If your data already exists in a SQL database, CSV file, or other structured format, LlamaIndex can query the data in these sources. This includes **text-to-SQL** (natural language to SQL operations) and also **text-to-Pandas** (natural language to Pandas operations).

- Text-to-SQL Guide
- Text-to-Pandas Guide

>> Ask AI

Advanced QA Topics

As you scale to more complex questions / more data, there are many techniques in LlamaIndex to help you with better query understanding, retrieval, and integration of data sources.

- **Querying Complex Documents**: Oftentimes your document representation is complex - your PDF may have text, tables, charts, images, headers/footers, and more. LlamaIndex provides advanced indexing/retrieval integrated with LlamaParse, our proprietary document parser. Full cookbooks here.

- **Combine multiple sources**: is some of your data in Slack, some in PDFs, some in unstructured text? LlamaIndex can combine queries across an arbitrary number of sources and combine them.

    - Example of combining multiple sources

- **Route across multiple sources**: given multiple data sources, your application can first pick the best source and then "route" the question to that source.

    - Example of routing across multiple sources

- **Multi-document queries**: some questions have partial answers in multiple data sources which need to be questioned separately before they can be combined

    - Example of multi-document queries

    - Building a multi-document agent over the LlamaIndex docs - Text to SQL

Resources

LlamaIndex has a lot of resources around QA / RAG. Here are some core resource guides to refer to.

I'm a RAG beginner and want to learn the basics: Take a look at our "Learn" series of guides.

I've built RAG, and now I want to optimize it: Take a look at our "Advanced Topics" Guides.

I'm more advanced and want to build a custom RAG workflow: Use LlamaIndex workflows to compose advanced, agentic RAG pipelines, like this Corrective RAG workflow.

I want to learn all about a particular module: Here are the core module guides to help build simple-to-advanced QA/RAG systems:

- Query Engines

- Chat Engines

- Agents

⏩ Ask AI

Further examples

For further examples of Q&A use cases, see our Q&A section in Putting it All Together.

Ask AI