Refine

The refine documents chain constructs a response by looping over the input documents and iteratively updating its answer. For each document, it passes all non-document inputs, the current document, and the latest intermediate answer to an LLM chain to get a new answer.

Since the Refine chain only passes a single document to the LLM at a time, it is well-suited for tasks that require analyzing more documents than can fit in the model's context. The obvious tradeoff is that this chain will make far more LLM calls than, for example, the Stuff documents chain. There are also certain tasks which are difficult to accomplish iteratively. For example, the Refine chain can perform poorly when documents frequently cross-reference one another or when a task requires detailed information from many documents.

Refine documents chain

for i in len(docs)

"What is foo?"

"foo.."

docs[i]

Prompt

Question: {q}
Intermediate answer:
{answer[i-1]}
Context: {docs[i]}

"Question:
  What is foo?
Intermediate
answer:
  foo is...
Doc:
  foo is also..."

Model

"Foo is ..."