Enhancing RAG with Fast GraphRAG and InstructLab : A Scalable, Interpretable, and Efficient Framework

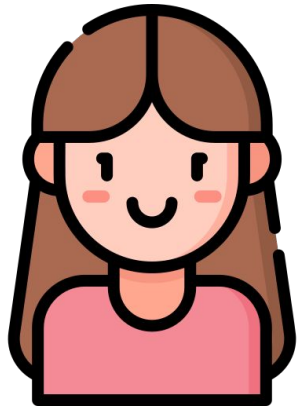
Tuhin Sharma
Technical Advisor, Data & AI @Red Hat

Red Hat
Data and AI Team

Agenda

- Introduction
- Set the Framework
- Zero shot LLM
- Classic RAG, Demo, Benefits, Drawback
- Microsoft GraphRAG, Demo, Benefits, Drawback
- HippoRAG, Demo, Benefits, Drawback
- Fast GraphRAG, Demo, Benefits, Drawback
- Comparative Analysis
- Supercharge Fast GraphRAG with InstructLab and LlamaIndex
- References

Make the books talk



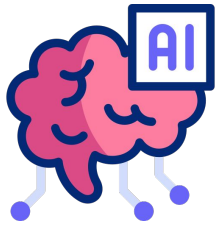
Maya,
Data Scientist

Set the Framework

Christmas Carol
By Charles Dickens

GPT-3.5-turbo

text-embedding-3-small



Data Generator
Claude 3.7 Sonnet

Create a set of 100 questions which are extremely complex to answer for a RAG system, based on the following text. {book} Make sure only sophisticated GRAPHRAG systems can only answer these correctly and the answers does not exceed 300 characters. Also provide the respective correct answers for the questions. Create the dataset in tabular format.

Evaluator
GPT-4

You are an expert evaluator.

Question: {question}
Golden Answer: {golden}
Model Answer: {prediction}

Evaluate the model answer against the golden answer. Respond with a score between 1 (poor) and 5 (perfect) based on accuracy, relevance, and completeness.

Question	**How does Dickens establish Scrooge's character through environmental imagery rather than direct description? Make sure the answer does not exceed 300 characters.**
Golden Answer	**Through cold imagery: he "iced his office," carries "his own low temperature," and is compared to "flint" with no "generous fire." The external cold reflects his internal emotional frigidity.**

Ask LLM directly (Zero Shot)

```python
from llama_index.llms.openai import OpenAI

llm = OpenAI(model="gpt-3.5-turbo")

query = """How does Dickens establish Scrooge's character through
environmental imagery rather than direct description? Make sure the
answer does not exceed 300 characters."""

response = llm.complete(query)
print(str(response))
```

```
Dickens uses cold, dark, and dreary settings to reflect Scrooge's personality.
```

3.2

Question

How does Dickens establish Scrooge's character through environmental imagery rather than direct description? Make sure the answer does not exceed 300 characters.

Golden Answer

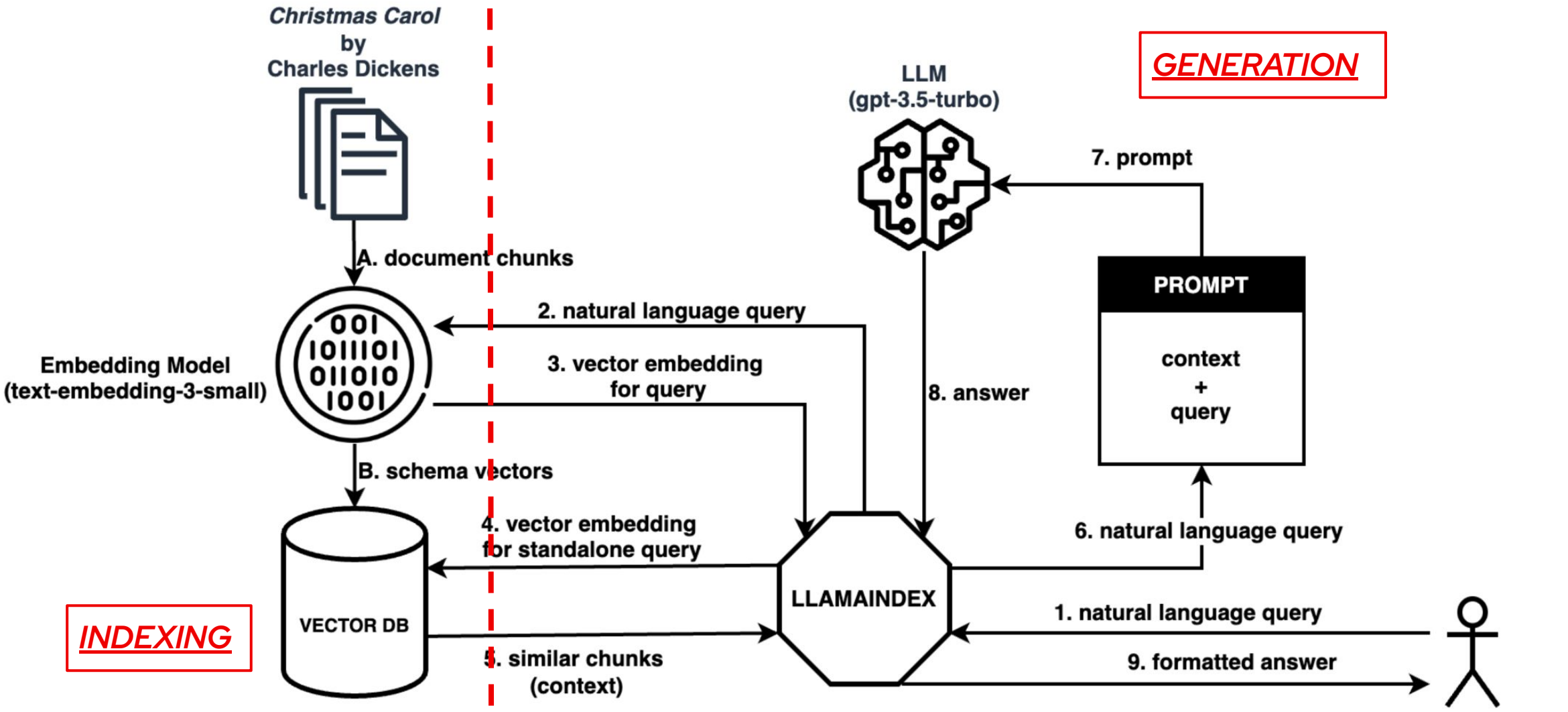
Through cold imagery: he "iced his office," carries "his own low temperature," and is compared to "flint" with no "generous fire." The external cold reflects his internal emotional frigidity.

Answer

Dickens uses cold, dark, and dreary settings to reflect Scrooge's personality.

https://github.com/tuhinsharma121/ai-playground/rag/simple-rag/llamaindex-rag/simple-rag-christmas.ipynb

Enter RAG : A New Hope



Red Hat
Data and AI Team

Simple RAG: Code (Indexing)

```python
import chromadb
from llama_index.vector_stores.chroma import ChromaVectorStore
from llama_index.core import VectorStoreIndex
from llama_index.core.node_parser import SentenceSplitter
from llama_index.embeddings.openai import OpenAIEmbedding

from llama_index.core.node_parser import SentenceSplitter
from llama_index.core import Document

# Load or create your document

with open("./book.txt") as f:
    doc = f.read()
text = doc


document = Document(text=text)

# Initialize Chroma client
chroma_client = chromadb.EphemeralClient()

# Create a collection for storing vectors
chroma_collection = chroma_client.get_or_create_collection("book_collection")

# Create the vector store
vector_store = ChromaVectorStore(chroma_collection=chroma_collection)

from llama_index.core import StorageContext

# Initialize the storage context
storage_context = StorageContext.from_defaults(vector_store=vector_store)

embed_model = OpenAIEmbedding(model="text-embedding-3-small")

# Create a sentence splitter for chunking text
parser = SentenceSplitter(chunk_size=1024, chunk_overlap=20)

# Build the index
index = VectorStoreIndex.from_documents([document], storage_context=storage_context,
                                        transformations=[parser], show_progress=True)
```

https://github.com/tuhinsharma121/ai-playground/rag/simple-rag/llamaindex-rag/simple-rag-christmas.ipynb

Simple RAG: Code (Generation)

```python
from llama_index.core.retrievers import VectorIndexRetriever
from llama_index.core.query_engine import RetrieverQueryEngine
from llama_index.llms.openai import OpenAI

retriever = VectorIndexRetriever(index, similarity_top_k=3, filter=None)
llm = OpenAI(model="gpt-3.5-turbo")
query_engine = RetrieverQueryEngine.from_args(retriever, llm=llm)

from llama_index.core import PromptTemplate

new_prompt_template_str = (
    "Context information is below.\n"
    "---------------------\n"
    "{context_str}\n"
    "---------------------\n"
    "Given the context and not prior knowledge, "
    "answer the query in less than 15 words.\n"
    "Query: {query_str}\n"
    "Answer: "
)

new_prompt_template = PromptTemplate(new_prompt_template_str)
query_engine.update_prompts({"response_synthesizer:text_qa_template": new_prompt_template})

query = """How does Dickens establish Scrooge's character through
environmental imagery rather than direct description?
Make sure the answer does not exceed 300 characters."""

response = query_engine.query(query)
print(str(response))
```

```
Dickens uses settings like a bleak moor and a desolate lighthouse to reflect Scrooge's cold and isolated personality.
```

3.5

Question

How does Dickens establish Scrooge's character through environmental imagery rather than direct description? Make sure the answer does not exceed 300 characters.

Golden Answer

Through cold imagery: he "iced his office," carries "his own low temperature," and is compared to "flint" with no "generous fire." The external cold reflects his internal emotional frigidity.

Answer

Dickens establishes Scrooge's character through bleak, desolate settings and contrasting joyful scenes.

https://github.com/tuhinsharma121/ai-playground/rag/simple-rag/llamaindex-rag/simple-rag-christmas.ipynb

Simple RAG : Performance (Golden Data of 100 QnA)

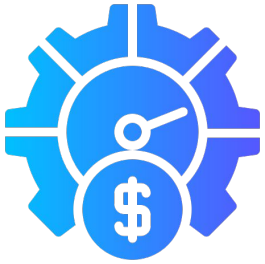
RAG Type	INDEXING						
	time	cost	gpt-3.5-turbo			text-embedding-3-small	
			#input token	#output token	#request	#input token	#request
simple-rag	4s	$ 0.01	-	-	-	39k	1

RAG Type	Eval Score
simple-rag	3.68

RAG Type	GENERATION						
	time	cost	gpt-3.5-turbo			text-embedding-3-small	
			#input token	#output token	#request	#input token	#request
simple-rag	2 min	$ 0.15	283k	1.6k	100	41k	100

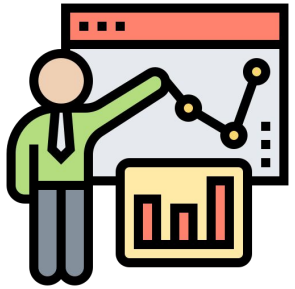
Simple RAG : Pros & Cons

ADVANTAGES



Lower Cost & Faster Time to Value

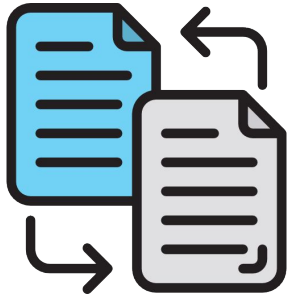
Up-to-Date Knowledge Injection



Easy to Interpret and Debug

DISADVANTAGES

Limited context window

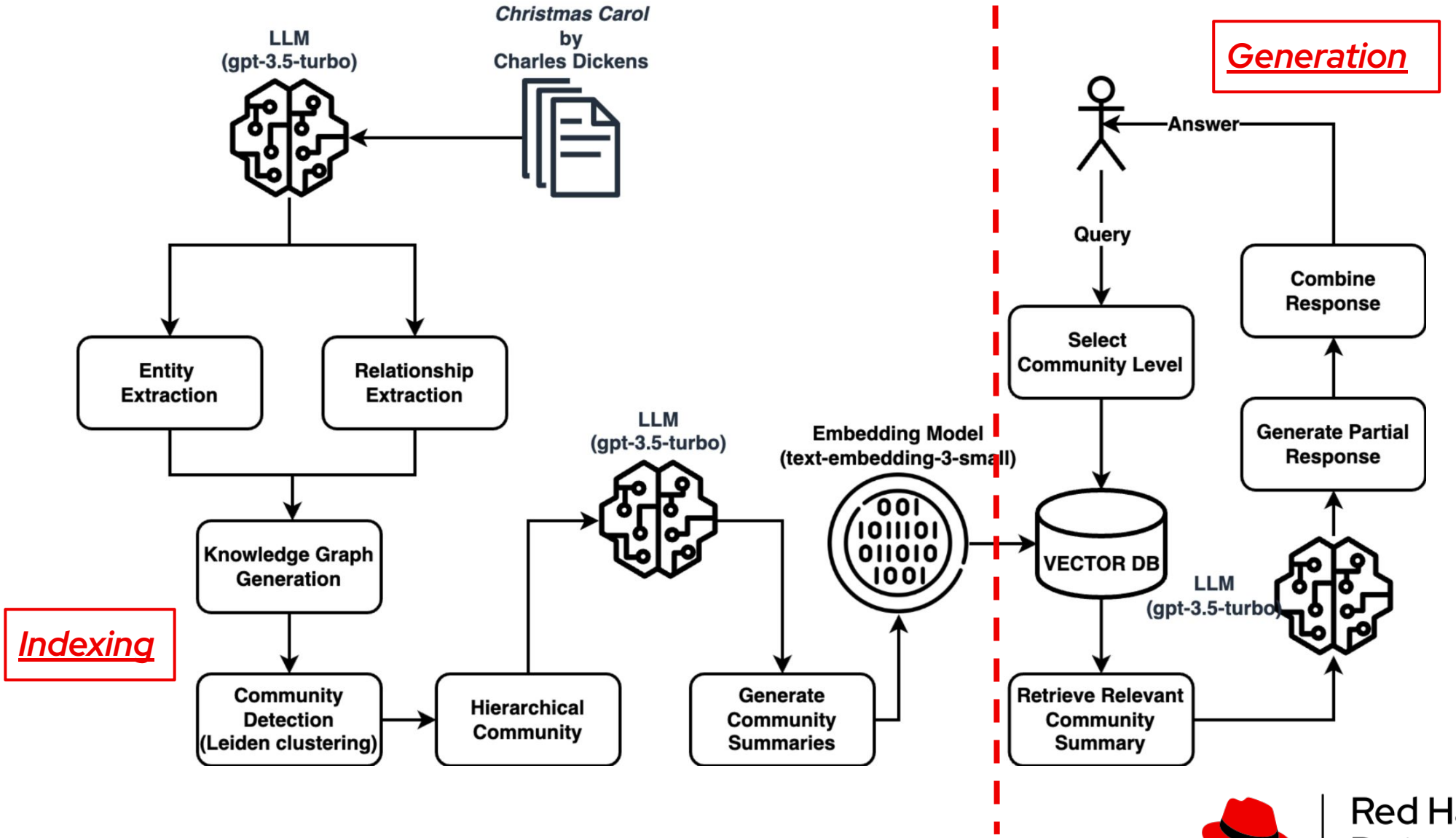


No Relationships between documents

Complex reasoning questions

Red Hat
Data and AI Team

Microsoft GraphRAG : Connecting the Dots



From Local to Global: A GraphRAG Approach to Query-Focused Summarization

Microsoft RAG: Code (Indexing)

```
python -m graphrag init --root ./book_index
```

Initialize root folder for Indexing

```
cp book.txt ./book_index/input/book.txt
```

Copy book.txt in the input folder

```
models:
  default_chat_model:
    api_key: ${OPENAI_API_KEY}
    model: gpt-3.5-turbo
default_embedding_model:
    api_key: ${OPENAI_API_KEY}
    model: text-embedding-3-small
snapshots:
  graphml: true
chunks:
  size: 1024
  overlap: 20
```

Initialize variable in settings.yaml

```
python -m graphrag index --root ./book_index
```

Create graphrag index

https://github.com/microsoft/graphrag
https://github.com/tuhinsharma121/ai-playground/blob/master/rag/graphrag/ms-graphrag/ms-graphrag-global.ipynb

Microsoft RAG: Code (Generation)

```
import subprocess
from IPython.display import Markdown
import os
import re


def ask_graph(query):
    command = ['/opt/anaconda3/envs/ms-graphrag-dev/bin/python',
               '-m',
               'graphrag',
               'query',
               '--root',
               './book_index',
               '--method',
               'global',
               '--query',
               query]
    output = subprocess.check_output(command,universal_newlines=True,stderr=subprocess.STDOUT)
    text = output.split("Search Response:\n")[1]
    return re.sub(r'\[Data:.*?\]','',text).strip()

query = """How does Dickens establish Scrooge's character through
environmental imagery rather than direct description?
Make sure the answer does not exceed 300 characters."""

ask_graph(query)
```

```
"**Dickens portrays Scrooge's character through environmental imagery, like his cold, dark office
contrasting with warm Christmas settings, reflecting his isolation and lack of empathy . This indi
rect approach reveals Scrooge's inner traits through the external world he inhabits, adding depth
to his characterization.**"
```

4.5

Question

How does Dickens establish Scrooge's character through environmental imagery rather than direct description? Make sure the answer does not exceed 300 characters.

Golden Answer

Through cold imagery: he "iced his office," carries "his own low temperature," and is compared to "flint" with no "generous fire." The external cold reflects his internal emotional frigidity.

Answer

Dickens' Characterization of Scrooge\n\nDickens skillfully portrays Scrooge's character through vivid environmental imagery. The somber settings of Scrooge's chambers, with details like damp walls and broken windows, symbolize his isolation and miserly nature . The cold, dark atmosphere of his office reflects his cold and miserly personality, emphasizing his lack of compassion .**

https://github.com/microsoft/graphrag
https://github.com/tuhinsharma121/ai-playground/blob/master/rag/graphrag/ms-graphrag/ms-graphrag-global.ipynb

Microsoft GraphRAG : Performance (Golden Data of 100 QnA)

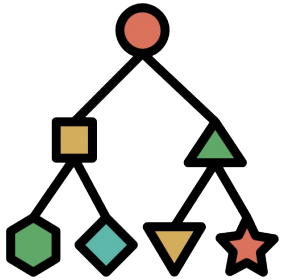
RAG Type	Indexing						
	time	cost	gpt-3.5-turbo			text-embedding-3-small	
			#input token	#output token	#request	#input token	#request
simple-rag	4s	$ 0.01	-	-	-	39k	1
ms-graphrag	**2 min 20 sec**	**$ 0.31**	**377k**	**78k**	**224**	**68k**	**33**

RAG Type	Answer Generation						
	time	cost	gpt-3.5-turbo			text-embedding-3-small	
			#input token	#output token	#request	#input token	#request
simple-rag	2 min	$ 0.15	283k	1.6k	100	41k	100
ms-graphrag	**24 min 30 sec**	**$ 1.00**	**2.3M**	**90k**	**294**	**-**	**-**

RAG Type	Eval Score
simple-rag	3.68
ms-graphrag	**4.42**

Microsoft GraphRAG : Pros & Cons

ADVANTAGES



Map Reduce to tackle Limited Context Window

Captures semantic relationships between information pieces

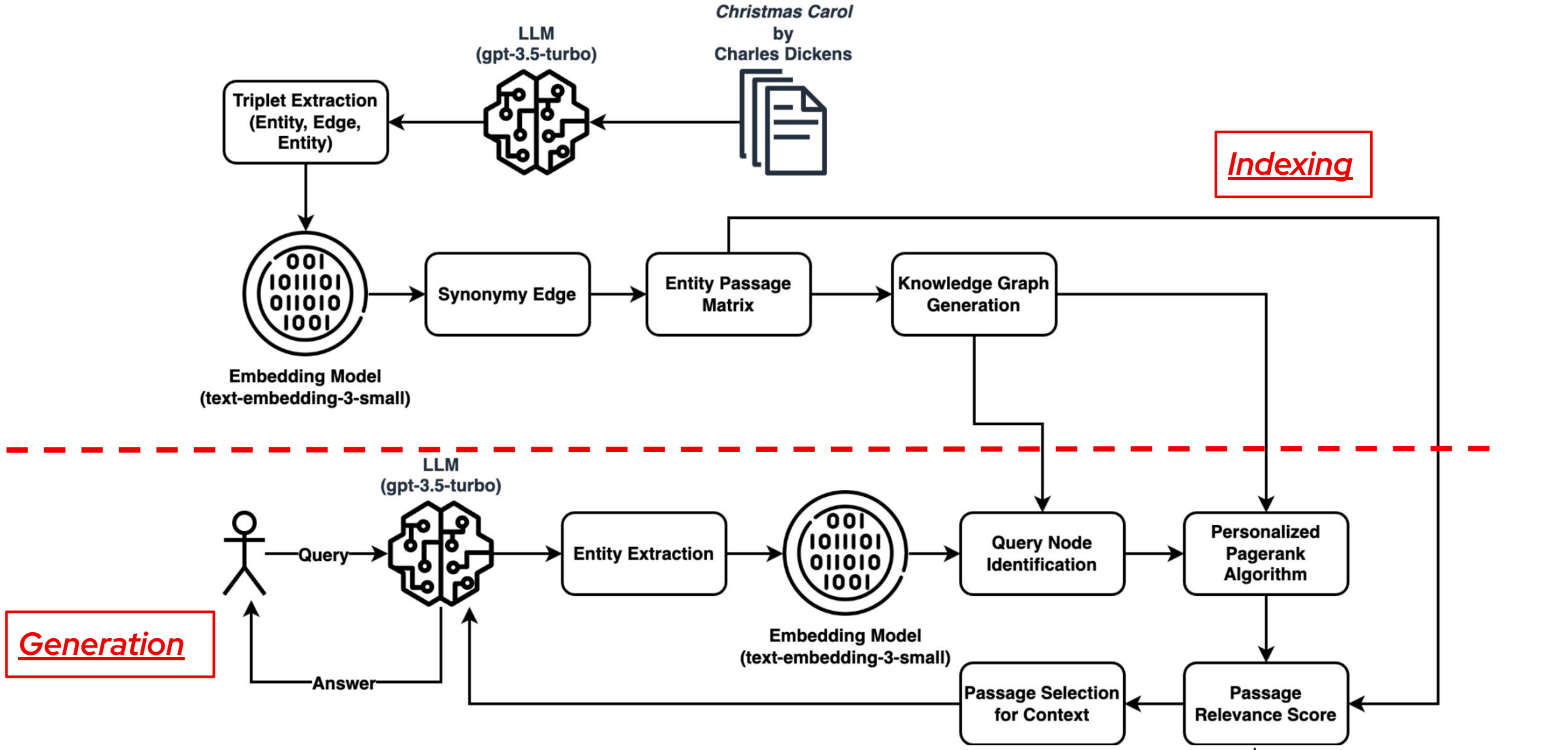
Handles complex hierarchical information better

DISADVANTAGES

Graph construction expensive

Query latency increases with graph size

HippoRAG : Memory Matters



HippoRAG: Neurobiologically Inspired Long-Term Memory for Large Language Models

HippoRAG : Code (Data Preparation)

```python
import nest_asyncio
import asyncio
nest_asyncio.apply()
from llama_index.core.node_parser import SentenceSplitter
from llama_index.core import Document

# Load or create your document
with open("./book.txt") as f:
    doc = f.read()
document = Document(text=doc)

# Initialize the splitter
splitter = SentenceSplitter(
    chunk_size=1024,    # Maximum number of characters per chunk
    chunk_overlap=20,   # Number of characters overlapping between chunks
)

# Parse the document into sentence-level nodes
nodes = splitter.get_nodes_from_documents([document])

docs = list()
# Each node contains a sentence
for node in nodes:
    docs.append(node.text)
```


https://github.com/OSU-NLP-Group/HippoRAG
https://github.com/tuhinsharma121/ai-playground/rag/graphrag/hippo-graphrag/hippo-graphrag-global.ipynb

Red Hat
Data and AI Team

HippoRAG : Code (Indexing)

```python
from hipporag import HippoRAG

save_dir = 'hipporag_books'
llm_model_name = 'gpt-3.5-turbo'
embedding_model_name = 'text-embedding-3-small'

#Startup a HippoRAG instance
hipporag = HippoRAG(save_dir=save_dir,
                    llm_model_name=llm_model_name,
                    embedding_model_name=embedding_model_name)

#Run indexing
hipporag.index(docs=docs)
```


https://github.com/OSU-NLP-Group/HippoRAG
https://github.com/tuhinsharma121/ai-playground/rag/graphrag/hippo-graphrag/hippo-graphrag-global.ipynb

HippoRAG : Code (Generation)

```
query = """How does Dickens establish Scrooge's character through
environmental imagery rather than direct description?
Make sure the answer does not exceed 300 characters."""

print(hipporag.rag_qa(queries=[query])[0][0].answer)
```

```
"Dickens establishes Scrooge's character through environmental imagery by desc
ribing the cold, dark, and gloomy setting around him, reflecting his own cold
and miserly personality. The fog, darkness, and icy weather mirror Scrooge's o
wn demeanor and lack of warmth towards others."
```

4.6

Question

How does Dickens establish Scrooge's character through environmental imagery rather than direct description? Make sure the answer does not exceed 300 characters.

Golden Answer

Through cold imagery: he "iced his office," carries "his own low temperature," and is compared to "flint" with no "generous fire." The external cold reflects his internal emotional frigidity.

Answer

Dickens establishes Scrooge's character through environmental imagery by describing the cold, dark, and gloomy setting around him, reflecting his own cold and miserly personality. The fog, darkness, and icy weather mirror Scrooge's own demeanor and lack of warmth towards others.

https://github.com/OSU-NLP-Group/HippoRAG
https://github.com/tuhinsharma121/ai-playground/rag/graphrag/hippo-graphrag/hippo-graphrag-global.ipynb

HippoRAG : Performance (Golden Data of 100 QnA)

RAG Type	Indexing						
	time	cost	gpt-3.5-turbo			text-embedding-3-small	
			#input token	#output token	#request	#input token	#request
simple-rag	4s	$ 0.01	-	-	-	39k	1
ms-graphrag	2 min 20 sec	$ 0.31	377k	78k	224	68k	33
hippo-rag	**2 min 15 sec**	**$ 0.09**	**109k**	**19k**	**90**	**50k**	**76**

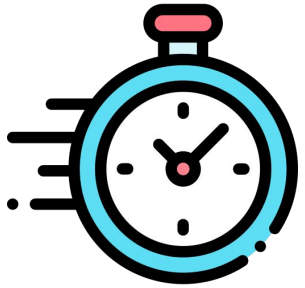
RAG Type	Answer Generation						
	time	cost	gpt-3.5-turbo			text-embedding-3-small	
			#input token	#output token	#request	#input token	#request
simple-rag	2 min	$ 0.15	283k	1.6k	100	41k	100
ms-graphrag	24 min 30 sec	$ 1.00	2.3M	90k	294	-	-
hippo-rag	**9 min 32 sec**	**$ 0.44**	**817k**	**21k**	**196**	**5k**	**197**

RAG Type	Eval Score
simple-rag	3.68
ms-graphrag	4.42
hippo-rag	**3.79**

HippoRAG : Pros & Cons

ADVANTAGES

Scales well incrementally



Faster query response

DISADVANTAGES

Accuracy degrades

Fast GraphRAG : Acceleration Breakthrough

Microsoft GraphRAG

HippoRAG

Red Hat
Data and AI Team

Fast GraphRAG : Code

```python
import nest_asyncio
import asyncio
nest_asyncio.apply()
from llama_index.core.node_parser import SentenceSplitter
from llama_index.core import Document

# Load or create your document
with open("./book.txt") as f:
    doc = f.read()
document = Document(text=doc)

# Initialize the splitter
splitter = SentenceSplitter(
    chunk_size=1024,    # Maximum number of characters per chunk
    chunk_overlap=20,   # Number of characters overlapping between chunks
)

# Parse the document into sentence-level nodes
nodes = splitter.get_nodes_from_documents([document])

docs = list()
# Each node contains a sentence
for node in nodes:
    docs.append(node.text)
```


https://github.com/circlemind-ai/fast-graphrag
https://github.com/tuhinsharma121/ai-playground/rag/graphrag/fast-graphrag/fast-graphrag-global.ipynb

Fast GraphRAG : Code

```python
from typing import List
import instructor
from dotenv import load_dotenv
from fast_graphrag import GraphRAG
from fast_graphrag._llm import OpenAIEmbeddingService, OpenAILLMService


DOMAIN = "Analyze this story and identify the characters. \
Focus on how they interact with each other, \
the locations they explore, and their relationships."

QUERIES = [
    "What is the significance of Christmas Eve in A Christmas Carol?",
    "How does the setting of Victorian London contribute to the story's themes?",
    "Describe the chain of events that leads to Scrooge's transformation.",
    "How does Dickens use the different spirits (Past, Present, and Future) to guide Scrooge?",
    "Why does Dickens choose to divide the story into \"staves\" rather than chapters?"
]

ENTITY_TYPES = ["Character", "Animal", "Place", "Object", "Activity", "Event"]
save_dir = 'fastgraphrag_books'
llm_model_name = 'gpt-3.5-turbo'
embedding_model_name = 'text-embedding-3-small'

fast_grag = GraphRAG(
    working_dir=save_dir,
    domain=DOMAIN,
    example_queries="\n".join(QUERIES),
    entity_types=ENTITY_TYPES,
    config=GraphRAG.Config(
        llm_service=OpenAILLMService(
            model=llm_model_name,
        ),
        embedding_service=OpenAIEmbeddingService(
            model=embedding_model_name,
        ),
    ),
)
for doc in docs:
    fast_grag.insert(doc)
```


https://github.com/circlemind-ai/fast-graphrag
https://github.com/tuhinsharma121/ai-playground/rag/graphrag/fast-graphrag/fast-graphrag-global.ipynb

Red Hat
Data and AI Team

Fast GraphRAG : Code

```
query = """How does Dickens establish Scrooge's character through
environmental imagery rather than direct description?
Make sure the answer does not exceed 300 characters."""

print(fast_grag.query(query).response)
```

```
Dickens establishes Scrooge's character through the environmental imagery of a lone
ly, dark house, with no soul expressing kindness towards him, contrasting his unfee
ling nature with a warm, bustling family scene filled with laughter and joy.
```

4.8

Question

How does Dickens establish Scrooge's character through environmental imagery rather than direct description? Make sure the answer does not exceed 300 characters.

Golden Answer

Through cold imagery: he "iced his office," carries "his own low temperature," and is compared to "flint" with no "generous fire." The external cold reflects his internal emotional frigidity.

Answer

Dickens establishes Scrooge's character through environmental imagery by describing his gloomy mansion, neglected offices, and cold, barren rooms to reflect Scrooge's inner bitterness and isolation.

https://github.com/circlemind-ai/fast-graphrag
https://github.com/tuhinsharma121/ai-playground/rag/graphrag/fast-graphrag/fast-graphrag-global.ipynb

Performance Analysis (Golden Data of 100 QnA)

RAG Type	Indexing						
	time	cost	gpt-3.5-turbo			text-embedding-3-small	
			#input token	#output token	#request	#input token	#request
simple-rag	4s	$ 0.01	-	-	-	39k	1
ms-graphrag	2 min 20 sec	$ 0.31	377k	78k	224	68k	33
hippo-rag	2 min 15 sec	$ 0.09	109k	19k	90	50k	76
fast graphrag	**9 min 55 sec**	**$ 0.21**	**231k**	**57k**	**157**	**31k**	**48**

RAG Type	Answer Generation						
	time	cost	gpt-3.5-turbo			text-embedding-3-small	
			#input token	#output token	#request	#input token	#request
simple-rag	2 min	$ 0.15	283k	1.6k	100	41k	100
ms-graphrag	24 min 30 sec	$ 1.00	2.3M	90k	294	-	-
hippo-rag	9 min 32 sec	$ 0.44	817k	21k	196	5k	197
fast graphrag	**5 min 43 sec**	**$ 0.45**	**908k**	**10k**	**196**	**5k**	**100**

RAG Type	Eval Score
simple-rag	3.68
ms-graphrag	4.42
hippo-rag	3.79
fast graphrag	**4.02**

Comparative Analysis of GraphRAG Techniques

Feature	Microsoft GraphRAG	HippoRAG	Fast GraphRAG
Graph Construction	Community clustering + summarization	Synonymy augmentation + clustering	Flat, fast, lightweight graphs
Retrieval Core	Map-Reduce Summarization	Personalized PageRank	Personalized PageRank
Incremental Updates	Limited	Moderate	Fully incremental
Speed	Slow	Moderate	High-speed retrieval

Supercharge Fast GraphRAG with LlamaIndex and InstructLab

```python
from typing import List
import instructor
from dotenv import load_dotenv
from fast_graphrag import GraphRAG
from fast_graphrag._llm import OpenAIEmbeddingService, OpenAILLMService


DOMAIN = "Analyze this story and identify the characters. \
Focus on how they interact with each other, \
the locations they explore, and their relationships."

QUERIES = [
    "What is the significance of Christmas Eve in A Christmas Carol?",
    "How does the setting of Victorian London contribute to the story's themes?",
    "Describe the chain of events that leads to Scrooge's transformation.",
    "How does Dickens use the different spirits (Past, Present, and Future) to guide Scrooge?",
    "Why does Dickens choose to divide the story into \"staves\" rather than chapters?"
]

ENTITY_TYPES = ["Character", "Animal", "Place", "Object", "Activity", "Event"]
save_dir = 'fastgraphrag_books'
llm_model_name = 'gpt-3.5-turbo'
embedding_model_name = 'text-embedding-3-small'

fast_grag = GraphRAG(
    working_dir=save_dir,
    domain=DOMAIN,
    example_queries="\n".join(QUERIES),
    entity_types=ENTITY_TYPES,
    config=GraphRAG.Config(
        llm_service=OpenAILLMService(
            model=llm_model_name,
        ),
        embedding_service=OpenAIEmbeddingService(
            model=embedding_model_name,
        ),
    ),
)
for doc in docs:
    fast_grag.insert(doc)
```

LlamaIndex

https://github.com/instructlab/instructlab
https://github.com/run-llama/llama_index

Red Hat
Data and AI Team

References

- Building a Simple Yet Powerful RAG Pipeline with LlamaIndex: A Christmas Carol Example
- GraphRAG : Beyond RAG with LlamaIndex for Smarter, Structured Retrieval
- Microsoft GraphRAG: Transforming Unstructured Text into Explainable, Queryable Intelligence using Knowledge Graph-Enhanced RAG
- HippoRAG : Redefining AI Retrieval emulating the Hippocampus
- Fast GraphRAG Architecture: Speed, Scalability, and Structured Retrieval for GenAI

Deck - https://bit.ly/tuhinsharma121

Code - https://github.com/tuhinsharma121/ai-playground

Blogs - https://medium.com/@tuhinsharma121

Linkedin - https://www.linkedin.com/in/tuhinsharma121

Github - https://github.com/tuhinsharma121

Website - https://tuhinsharma.com

THANK YOU

Red Hat
Data and AI Team