Information Retrieval Sommersemester 2024

Semval204Task1
LLMs as Annotations

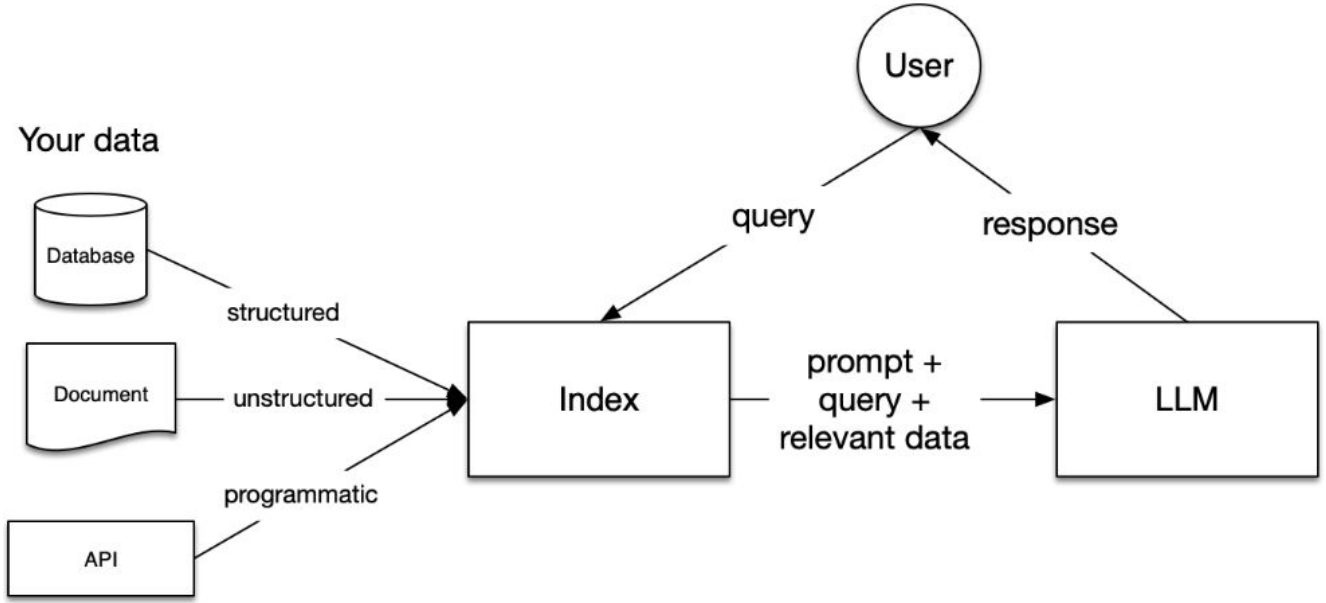
CIS-LMU
Sommersemester 2024
Dr. Robert Litschko

18.07.2024
Aldi Halili

Gliederung

- Einführung
- Daten
- LLM Sherpa Framework
- LlamaIndex Framework
- GPT Chat Model
- Evaluation

Question-Answering: Retrieval-Augmented Generation (RAG)



Ein RAG-System generiert Antworten basierend auf eigenen Dokumenten, ohne auf externe Informationen zuzugreifen.

LLMs sind auf umfangreichen Datensätzen trainiert, jedoch nicht speziell auf eigenen Daten.

Komponenten der Implementierung des Medizinischen RAG-Systems

- Datenquelle: Medizinische Fachbücher im PDF-Format
- PDF-Reader: LLM Sherpa Framework zur Textextraktion aus PDFs
- Indexierung: LlamaIndex Framework für die Organisation und das Retrieval von Daten
- Embeddings Model
- LLM Chat-Model: Einsatz von GPT-3.5 Turbo für die Generierung von Antworten
- Evaluierung: TruLens Framework zur Überprüfung und Bewertung der Modellleistung

Datenquellen: Medizinische Bücher

- Diabetes
  - Informationen über Diabetes Behandlungen und -management.
- General Pathology
  - grundlegende Einblicke in pathologische Prozesse und Krankheitsursachen.
- Neurology Introduction
  - Grundlage für die Erkennung und Behandlung neurologischer Krankheiten.
- The Practical Course in Clinical Medicine
  - Anwendungsbasierte Ansätze und klinische Verfahren für die medizinische Praxis.

Datenquelle-link

LLM Sherpa Framework - Fortschrittliches Tool zur PDF Textextraktion

- Überblick
  - Speziell für den Einsatz in Retrieval-Augmented Generation (RAG) Systemen konzipiert.
  - Open-Source Framework
  - PDF Rader: Klasse LayoutPDFReader
- Vorteile für das RAG-System:
  - Verbessert die Genauigkeit der Textextraktion, was zu präziseren Antworten des RAG-Systems führt.

Anwendung der Klasse “LayoutPDFReader”

```
# This step in using LayoutPDFReader to provide a url or file path to it and get back a document object.
from llmsherpa.readers import LayoutPDFReader

llmsherpa_api_url = "https://readers.llmsherpa.com/api/document/developer/parseDocument?renderFormat=all"
pdf_reader = LayoutPDFReader(llmsherpa_api_url)

# Access and read each file
for file_name in files:
    file_path = os.path.join(directory_path, file_name)
    if os.path.exists(file_path):
        print(f"Accessing file: {file_path}")
        doc = pdf_reader.read_pdf(file_path)
    else:
         print(f"File not found: {file_path}")
```

```
Accessing file: /content/drive/MyDrive/row_data/01. The Practical Course in Clinical Medicine Autor Władysław Grabski, Dariusz Nowak.pdf
Accessing file: /content/drive/MyDrive/row_data/1. Diabetes  Author Dr Mrs Anjali Kulkarni.pdf
Accessing file: /content/drive/MyDrive/row_data/1. General Pathology Author Mesele Bezabeh, Abiye Tesfaye.pdf
Accessing file: /content/drive/MyDrive/row_data/1. Neurology Introduction Author MUK Publications.pdf
```

```
print(type(doc))
```

```
<class 'llmsherpa.readers.layout_reader.Document'>
```

Funktionalitäten der LLM Sherpa

Texterkennung: Besonders effizient in der Extraktion von Listen, Tabellen und Abschnitten etc.

class
`llmsherpa.readers.layout_reader.`**`Document`**`(`*`blocks_json`*`)`

Bases: **`object`**

A document is a tree of blocks. It is the root node of the layout tree.

`chunks()`

Returns all the chunks in the document. Chunking automatically splits the document into paragraphs, lists, and tables without any prior knowledge of the document structure.

`sections()`

Returns all the sections in the document. This is useful for getting all the sections in a document.

`tables()`

Returns all the tables in the document. This is useful for getting all the tables in a document.

```
for chunk in doc.chunks():
    print(chunk.to_text())
```

Ausgeblendete Ausgabe anzeigen

```
print("Länge von chunks: ",len(doc.chunks()))
```

```
Länge von chunks:  1578
```

```
for table in doc.tables():
    print(table.to_text())
```

Ausgeblendete Ausgabe anzeigen

```
print("Es gibt insgesamt ",len(doc.tables()
```

```
Es gibt insgesamt  4 Tabellen
```

Llama Index Framework

- Open Source LlamaIndex: Ein Framework zur Entwicklung von kontext-augmentierten generativen KI-Anwendungen mit LLMs
- Gründung: November 2023
- Anwendung:
  - Question-Answering (RAG)
  - Chatbots,
  - Multi-modal Applikationen.
  - Fine tuning
  - Agenten

Implementierung von LlamaIndex Framework

```
# llama-index' Bibliothek installieren
!pip install llama-index
```

Ausgeblendete Ausgabe anzeigen

```
"""
VectorStoreIndex: Ein Index, der Vektorspeicherung nutzt, um relevante Informationen für die Anfrage zu finden.

ServiceContext: Hält verschiedene Service-Komponenten. z. B. Tokenizer oder Vektor-Embedding-Modelle.

StorageContext: Ermöglicht die Speicherung und Wiederherstellung von Indexen und Abfragen.

load_index_from_storage: Hilft dabei, einen bereits vorhandenen Index aus einem vorherigen Speicherstand zu laden.

Document: aus dem llama_index.core.readers-Modul ermöglicht die Erstellung, Bearbeitung und Verwaltung von Dokumenten
im LlamaIndex-Framework
"""
# benötigten Klassen von Lammaindex importieren
from llama_index.core import VectorStoreIndex, ServiceContext,StorageContext,load_index_from_storage
from llama_index.core.readers import Document
```

Open AI-API_Key, Chat Model, Prompt-Eingabe,

```python
# Sets the environment variable for the OpenAI API key
import os
os.environ["OPENAI_API_KEY"] = ""
```

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

h die Nutzung des **ServiceContext** und des definierten **System-Prompts** wird der Chatbot so konfiguriert, dass er ausschließlich rmationen auf Englisch aus den bereitgestellten Dokumenten generiert und keine externen Quellen verwendet.

```python
llm_context_query__service_context = ServiceContext.from_defaults(
llm=OpenAI(
model="gpt-3.5-turbo",
temperature=0.1,
),
system_prompt=
"""You are a friendly chatbot. Use exclusively only the following context to answer
the question at the end.Use only the available Information. If the answer is not from the context, then say "I have no information regarding yo

        Context: {context}

        Question: {question}

        Helpful answer:
""")
```

Embedding Model

- Open AI Embeddings Model
  - Verwendete Modell
- Local Model von Hugging-Face
- Embedding von GPT Promt-Eingabe

```python
from llama_index.embeddings.openai import OpenAIEmbedding
from llama_index.core import VectorStoreIndex
from llama_index.core import Settings

# global
Settings.embed_model = OpenAIEmbedding()

embed_model = Settings.embed_model


service_context = ServiceContext.from_defaults(embed_model=embed_model)
```

```python
!pip install llama-index-embeddings-huggingface



from llama_index.embeddings.huggingface import HuggingFaceEmbedding
from llama_index.core import Settings

Settings.embed_model = HuggingFaceEmbedding(
    model_name="BAAI/bge-small-en-v1.5"
)

embed_model = Settings.embed_model


service_context = ServiceContext.from_defaults(embed_model=embed_model)
```

Storing: Lokal speichern

VectorStoreIndex - wird 5 json Dateien herstellen:

default_vector_store
Datei speichert die Embeddings

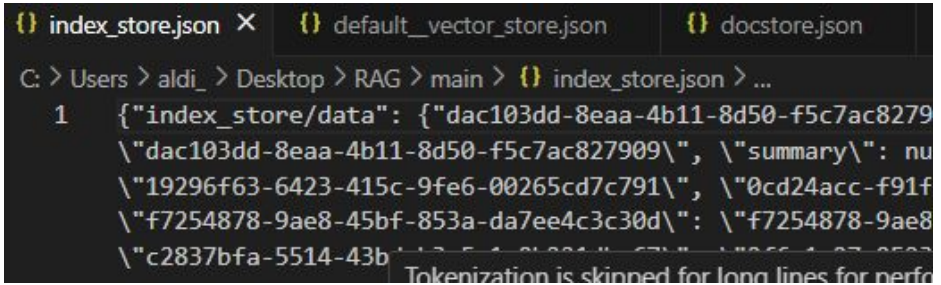
Index_store.json

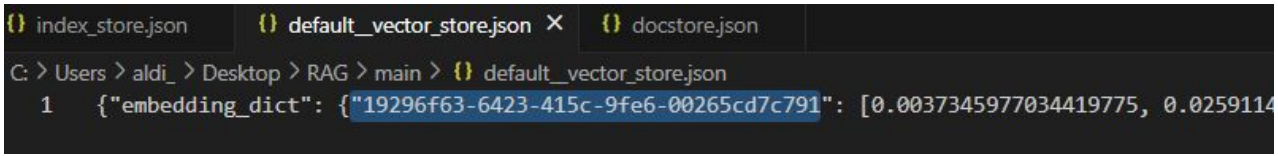
wird Indexierung gespeichert

Docstore

werden die Metadaten und text gespeichert.

```
PERSIST_DIR = "./storage"
if not os.path.exists(PERSIST_DIR):
    # Create a new index because the storage directory does not exist
    os.makedirs(PERSIST_DIR)  # Ensure the directory is created where the index will be stored
    index = index=VectorStoreIndex([],service_context=service_context)
    for chunk in doc.chunks():
      index.insert(Document(text=chunk.to_context_text(),extra_info={}))
    index.storage_context.persist(persist_dir=PERSIST_DIR)  # Persist the newly created index
else:
    # Load the existing index from the storage
    storage_context = StorageContext.from_defaults(persist_dir=PERSIST_DIR)
    index = load_index_from_storage(
        storage_context,
        service_context=service_context
    )
```





Antwort Generierung (Query Engine)

```
# Converts the index into a query engine, which can be used to perform queries.
query_engine=index.as_query_engine(service_contex=llm_context_query__service_context)
```

```
insulin1=query_engine.query("What is an insulin syringe and how can one handle it?")
print(insulin1)
```

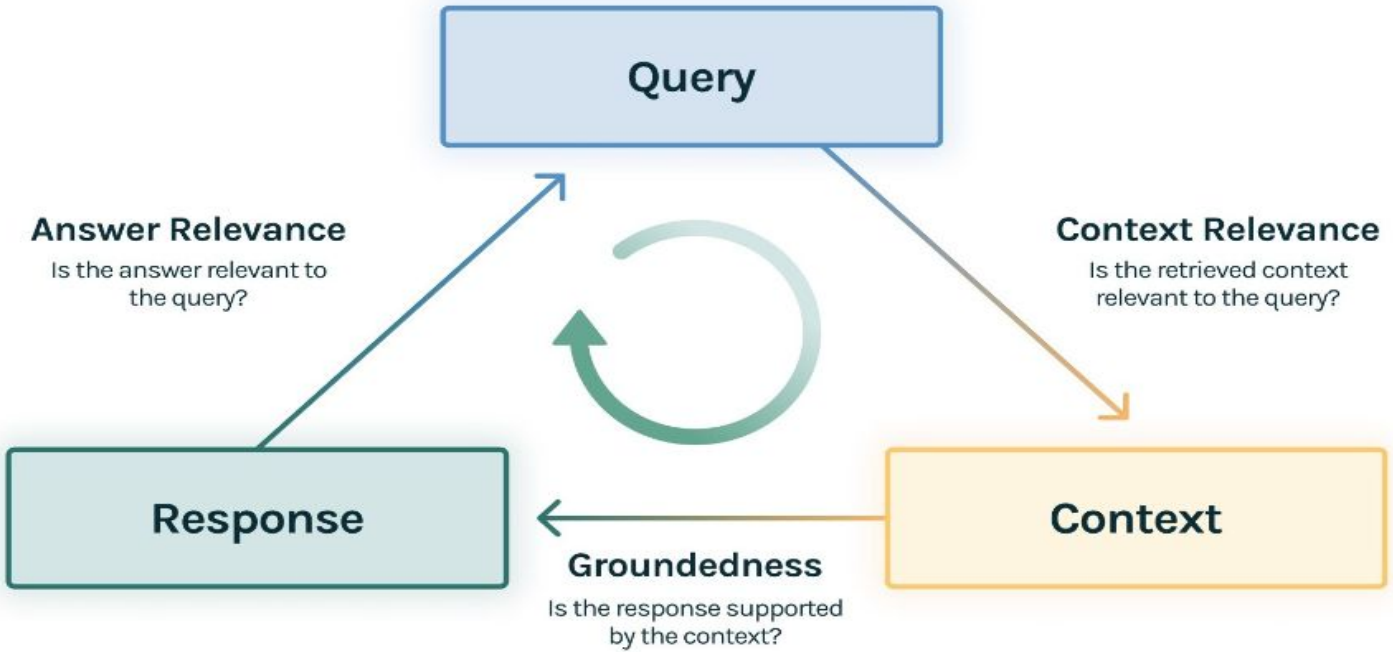
```
I do not have any information regarding your request.
```

```
insulin2=query_engine.query("What are insulin syringes?")
print(insulin2)
```

```
Insulin syringes are used to administer insulin for the treatment of diabetes.
```

Evaluation

Implementierung von RAG Triad Konzept anhand von Trulens Framework



Implementierung von TruLens

```
!pip install trulens_eval llama_index openai
```

```
#import from TruLens
from trulens_eval import Tru
tru = Tru()
```

- Trulens Bibliothek installieren und importieren von Trulens
- Feedback Funktionen initialisieren

```
from trulens_eval.feedback.provider import OpenAI
from trulens_eval import Feedback
import numpy as np

# Initialize provider class
provider = OpenAI()

# select context to be used in feedback. the location of context is app
from trulens_eval.app import App
context = App.select_context(query_engine)

# Define a groundedness feedback function
f_groundedness = (
    Feedback(provider.groundedness_measure_with_cot_reasons)
    .on(context.collect()) # collect context chunks into a list
    .on_output()
)

# Question/answer relevance between overall question and answer.
f_answer_relevance = (
    Feedback(provider.relevance)
    .on_input_output()
)
# Question/statement relevance between question and each context chunk.
f_context_relevance = (
    Feedback(provider.context_relevance_with_cot_reasons)
    .on_input()
    .on(context)
    .aggregate(np.mean)
)
```

Instrument app for logging with TruLens

```python
# Instrument app for logging with TruLens
from trulens_eval import TruLlama
tru_query_engine_recorder = TruLlama(query_engine,
    app_id='LlamaIndex_App1',
    feedbacks=[f_groundedness, f_answer_relevance, f_context_relevance])

# Liste der Fragen
question_q = [
    "What is a neurologist?",
    "What is a Treatment by Neurologist?",
    "What are Neurologists Tasks?",
```

```python
# Kontextmanager, um jede Frage einzeln zu stellen
with tru_query_engine_recorder as recording:
    for question in question_q:
        response = query_engine.query(question)
        print(f"Query: {question}")
        print(f"Response: {response}")
        print("\n")
```

Retrieve records and feedback

```python
import pandas as pd

# Retrieve records and feedback
records, feedback = tru.get_records_and_feedback(app_ids=["LlamaIndex_App1"])

# Convert records to a DataFrame
df_records = pd.DataFrame(records)
```

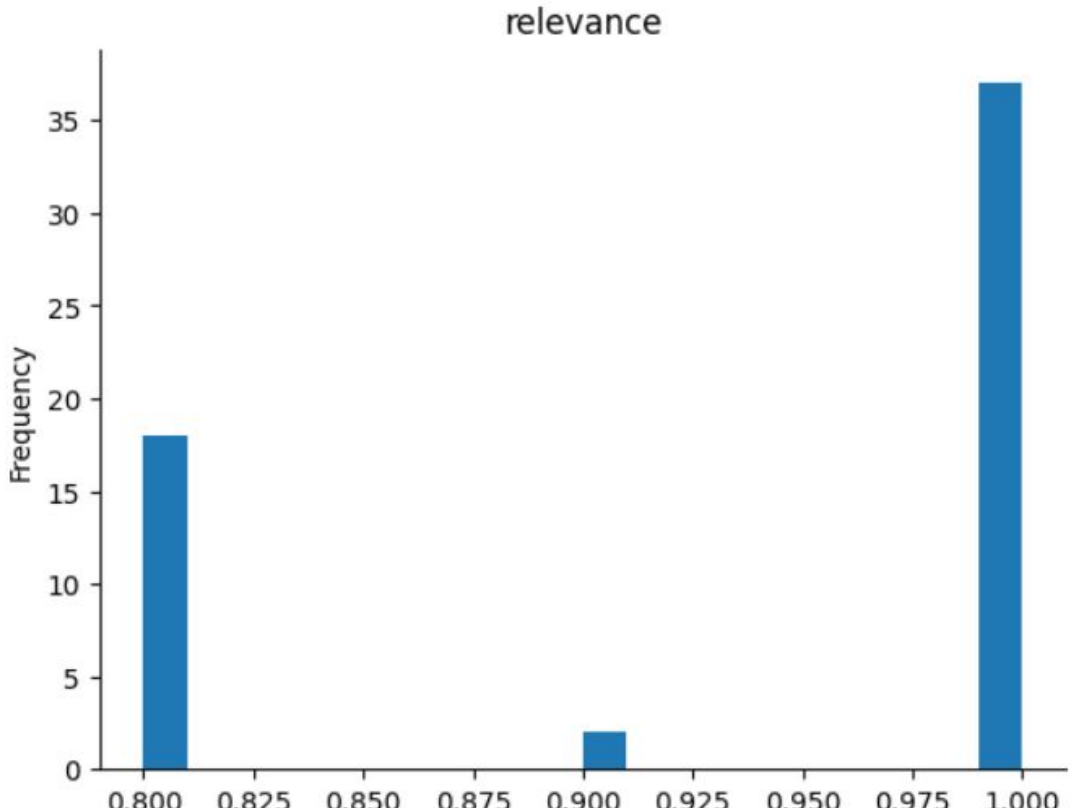
Average feedback values are returned and displayed in a range from 0 (worst) to 1 (best).

Bewertung

input	output	groundedness_measure_with_cot_reasons	context_relevance_with_cot_reasons	relevance
"What is a neurologist?"	"A neurologist is a medical doctor with specia...	1.000000	0.90	1.0
"What is a Treatment by Neurologist?"	"A Treatment by Neurologist typically involves...	1.000000	0.60	0.8
"What are Neurologists Tasks?"	"Neurologists' tasks include conducting resear...	1.000000	0.85	0.8

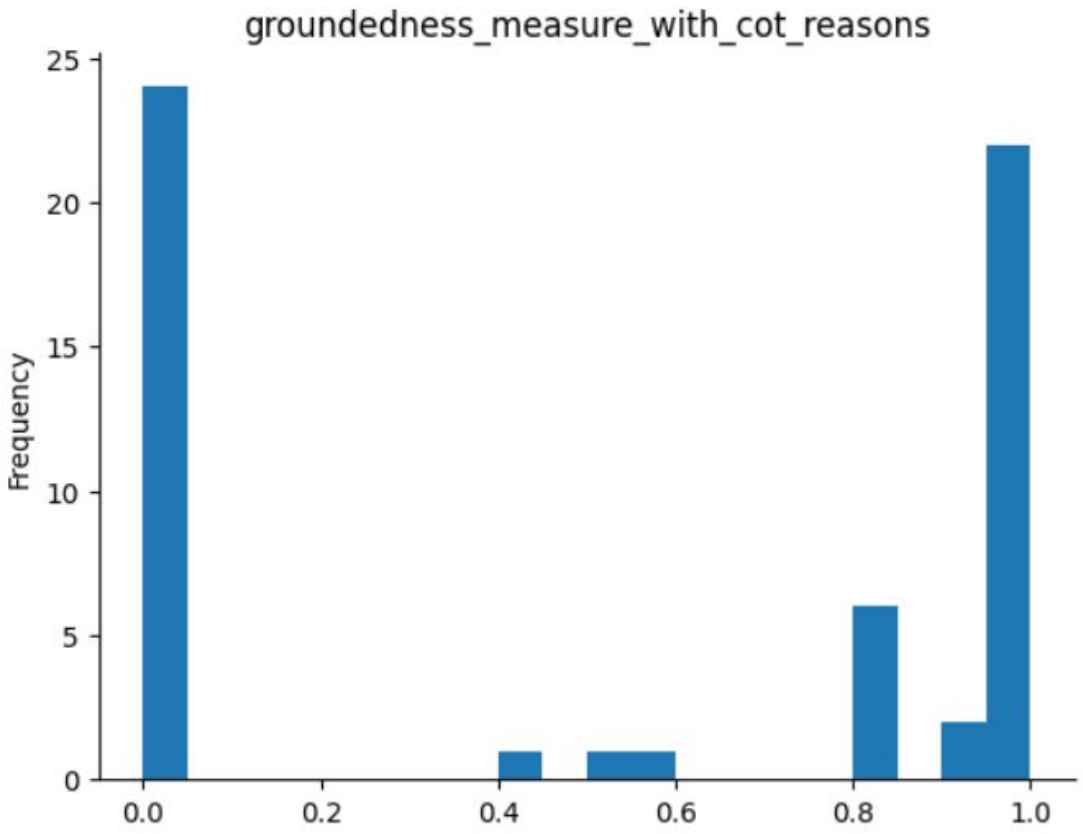
Question / Answer Relevance

Für 60 Anfragen liegt ein Score zwischen 08-1.0

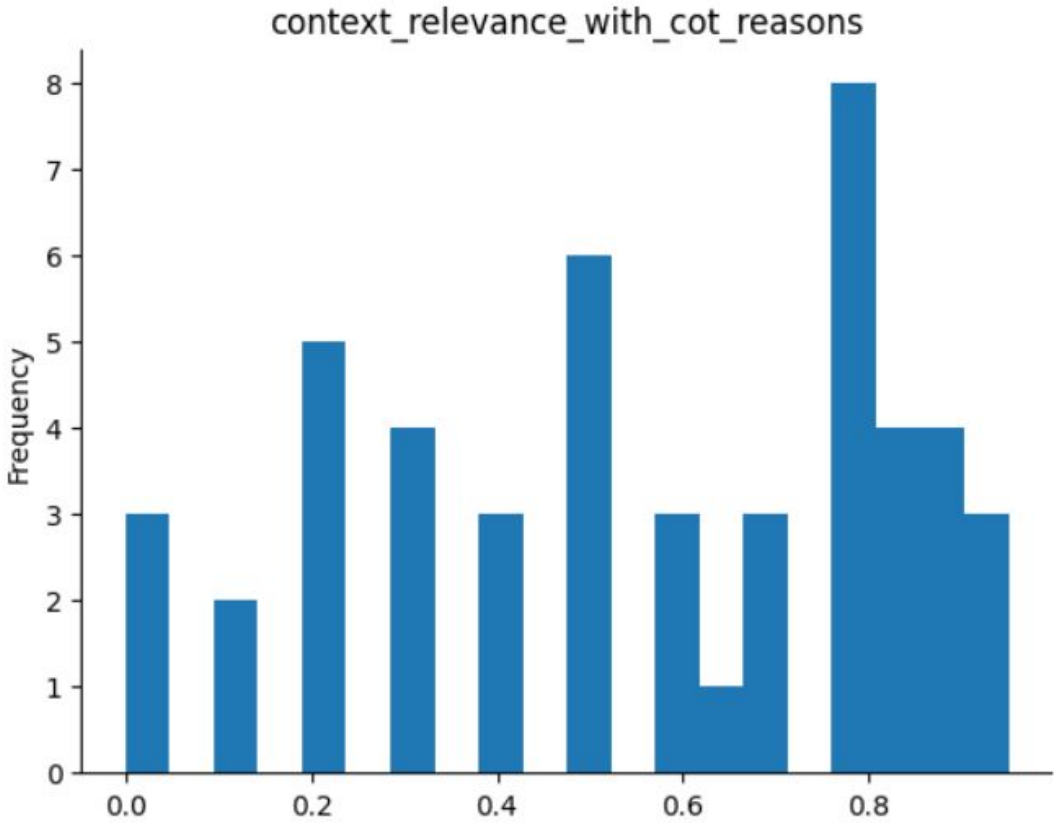


Groundedness

Does the answer rely on the context?



Question/statement relevance
between question and each context chunk.



Gesamtergebnis

```
tru.get_leaderboard(app_ids=["LlamaIndex_App1"])
```

app_id	groundedness_measure_with_cot_reasons	context_relevance_with_cot_reasons	relevance	latency	total_cost
LlamaIndex_App1	0.528509	0.559184	0.933333	2.210526	0.000547

Beispiele

index	input	output	groundedness_measure_with_cot_reasons	context_relevance_with_cot_reasons	relevance	groundedness_measure_with_cot_reasons_calls
0	"What is a neurologist?"	"A neurologist is a medical doctor with specialized training in diagnosing, treating, and managing disorders of the brain and nervous system."	1.0	0.9	1.0	[{'args': {'source': ['A doctor who has specialisation in neurology is known as a neurologist.\nThe neurologist treats disorders that affect the brain, spinal cord, and nerves, such as:', 'WHO'S NEUROLOGIST\nNeurologist is a medical doctor who possesses specialized training in diagnosing, treating and managing disorders of the brain and nervous system.\nPediatric neurologists are doctors with specialized training in children's neurological disorders.\nA neurologist's educational background and medical training includes an undergraduate degree, four years of medical school, a one-year internship and three years of specialized training.'], 'statement': 'A neurologist is a medical doctor with specialized training in diagnosing, treating, and managing disorders of the brain and nervous system.'}, 'ret': 1.0, 'meta': {'reasons': 'STATEMENT 0:\nCriteria: A neurologist is a medical doctor with specialized training in diagnosing, treating, and managing disorders of the brain and nervous system.\nSupporting Evidence: The source mentions that a neurologist is a medical doctor who possesses specialized training in diagnosing, treating, and managing disorders of the brain and nervous system.\nScore: 10\n'}}]
20	"What are insulin syringes?"	"I have no information regarding your request."	1.0	0.1	1.0	[{'args': {'source': ['Index > I\nImmunocytochemistry, 38, 284.', 'Progress > Indications of muscle disease\nGlycogen is a storage form of carbohydrate, and its breakdown is a source of energy.\nMuscle weakness is found in a rare group of hereditary diseases, the glycogen-storage diseases, in which various enzyme defects prevent the release of energy by the normal breakdown of glycogen in muscles.\nAs a result, abnormal amounts of glycogen are stored in the muscles and other organs.'], 'statement': 'I have no information regarding your request.'}, 'ret': 1.0, 'meta': {'reasons': 'STATEMENT 0:\nCriteria: Glycogen is a storage form of carbohydrate.\nMuscle weakness is found in a rare group of hereditary diseases, the glycogen-storage diseases.\nVarious enzyme defects prevent the release of energy by the normal breakdown of glycogen in muscles.\nAbnormal amounts of glycogen are stored in the muscles and other organs.\nSupporting Evidence: Progress > Indications of muscle disease. The source mentions that glycogen is a storage form of carbohydrate.\nScore: 10\n\nProgress > Indications of muscle disease. The source discusses how muscle weakness is found in a rare group of hereditary diseases known as glycogen-storage diseases.\nScore: 10\n\nProgress > Indications of muscle disease. The source explains that in glycogen-storage diseases, enzyme defects prevent the release of energy by the normal breakdown of glycogen in muscles.\nScore: 10\n\nProgress > Indications of muscle disease. The source states that in glycogen-storage diseases, abnormal amounts of glycogen are stored in the muscles and other organs.\nScore: 10\n'}}]

Weitere Beispiele von Dokumenten

	input	output	groundedness_measure_with_cot_reasons	context_relevance_with_cot_reasons	relevance	groundedness_measure_with_cot_reasons_calls
0	"What is a neurologist?"	"A neurologist is a medical doctor with specia...	1.000000	0.90	1.0	[{'args': {'source': ['A doctor who has specia...
1	"What is a Treatment by Neurologist?"	"A Treatment by Neurologist typically involves...	1.000000	0.60	0.8	[{'args': {'source': ['WHO'S NEUROLOGIST > Tre...
2	"What are Neurologists Tasks?"	"Neurologists' tasks include conducting resear...	1.000000	0.85	0.8	[{'args': {'source': ['NEUROLOGICAL EXAMINATIO...
3	"What is the Nervous System?"	"The nervous system is a control system of the...	1.000000	0.90	0.8	[{'args': {'source': ['ANATOMY AND FUNCTION OF...
4	"What is Neurophysiology?"	"Neurophysiology is a medical specialty that f...	1.000000	0.95	0.8	[{'args': {'source': ['Neurophysiology > Neuro...
5	"What are Nerve Cells?"	"Nerve cells are cells that come in different ...	1.000000	0.80	0.8	[{'args': {'source': ['Nerve Cells\nThere are ...
6	"What is an Axon?"	"The axon is a part of a neuron that carries n...	1.000000	0.90	0.8	[{'args': {'source': ['NEUROPHYSIOLOGY OF CENT...
7	"What is Pure alexia?"	"Pure alexia is the inability to recognize wor...	1.000000	0.90	0.9	[{'args': {'source': ['AGNOSIA > Agnosic Alexi...
8	"What is agnosia and what is the treatment abo...	"Agnosia is a condition where individuals have...	0.666667	NaN	0.8	[{'args': {'source': ['TREATMENT OF AGNOSIA\nF...
9	"What are the types of agnosia?"	"The types of agnosia are apperceptive and ass...	1.000000	0.90	1.0	[{'args': {'source': ['AGNOSIA > Forms of Agno...
10	"What is Alzheimer and what are the atypical A...	"I have no information regarding your request."	0.000000	0.80	1.0	[{'args': {'source': ['Alzheimer's disease\nAl...

Weitere Beispiele ohne Dokumente

40	"What is stomach carcinoma?"	"I have no information regarding your request."	1.00	0.20	1.0	[{'args': {'source': ['GENETIC TESTING > Retin...
41	"What is the general definition of carcinoma?"	"I have no information regarding your request."	1.00	0.40	1.0	[{'args': {'source': ['Meaning of the Results ...
42	"What causes endocarditis?"	"I have no information regarding your request."	1.00	0.30	1.0	[{'args': {'source': ['Slowly Progressive > Be...
43	"What is the difference between acute and chro...	"Acute kidney failure occurs suddenly and is u...	0.00	0.20	0.9	[{'args': {'source': ['Progress > Indications ...
44	"What are all the symptoms of meningitis?"	"I have no information regarding your request."	0.50	NaN	1.0	[{'args': {'source': ['Slowly Progressive > Be...
45	"Why are leukocytes and CRP levels in the bloo...	"Leukocytes and CRP levels in the blood are im...	0.00	0.80	0.8	[{'args': {'source': ['Laboratory tests\nRouti...
46	"When is an indication for a bypass surgery?"	"I have no information regarding your request."	0.00	NaN	1.0	[{'args': {'source': ['Vascular dementia > Beh...
47	"What are the main causes of climate change?"	"I have no information regarding your request."	0.00	NaN	1.0	[{'args': {'source': ['Etiology-pathogenesis o...
48	"How does blockchain technology work?"	"I have no information regarding your request."	0.00	NaN	1.0	[{'args': {'source': ['Reasons for ethical Con...
49	"What are the benefits and challenges of remot...	"Benefits of remote work include increased fle...	0.00	NaN	0.8	[{'args': {'source': ['Reasons for ethical Con...
50	"Can you explain the theory of relativity in s...	"I have no information regarding your request."	1.00	NaN	1.0	[{'args': {'source': ['NEURAL CORRELATES OF CO...

Literatur

Evaluate and Track your LLM Experiments: Introducing TruLens | by Shayak Sen | TruLens | Medium

Build and Evaluate LLM Apps with LlamaIndex and TruLens | by Shayak Sen | TruLens | Medium

📓 LlamaIndex Quickstart - 🦑 TruLens

Query Engine - LlamaIndex

Question-Answering (RAG) - LlamaIndex

Indexing & Embedding - LlamaIndex