UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

Dipartimento di
Informatica, Sistemistica
e Comunicazione

ReGAInS

UNIVERSITÀ
DI TORINO

Retrieval-Augmented Generation: Foundations, Evolution, and Applications to Privacy Risk and Misinformation Detection

Marco Viviani

University of Milano-Bicocca

Department of Informatics, Systems, and Communication (DISCo)

July 22, 2025

Background: IR + Text Generation

What is RAG?

- Integrating **Information Retrieval** (IR) Techniques in **Text Generation**

Information Retrieval

+

Text Generation

Close-book exam

=

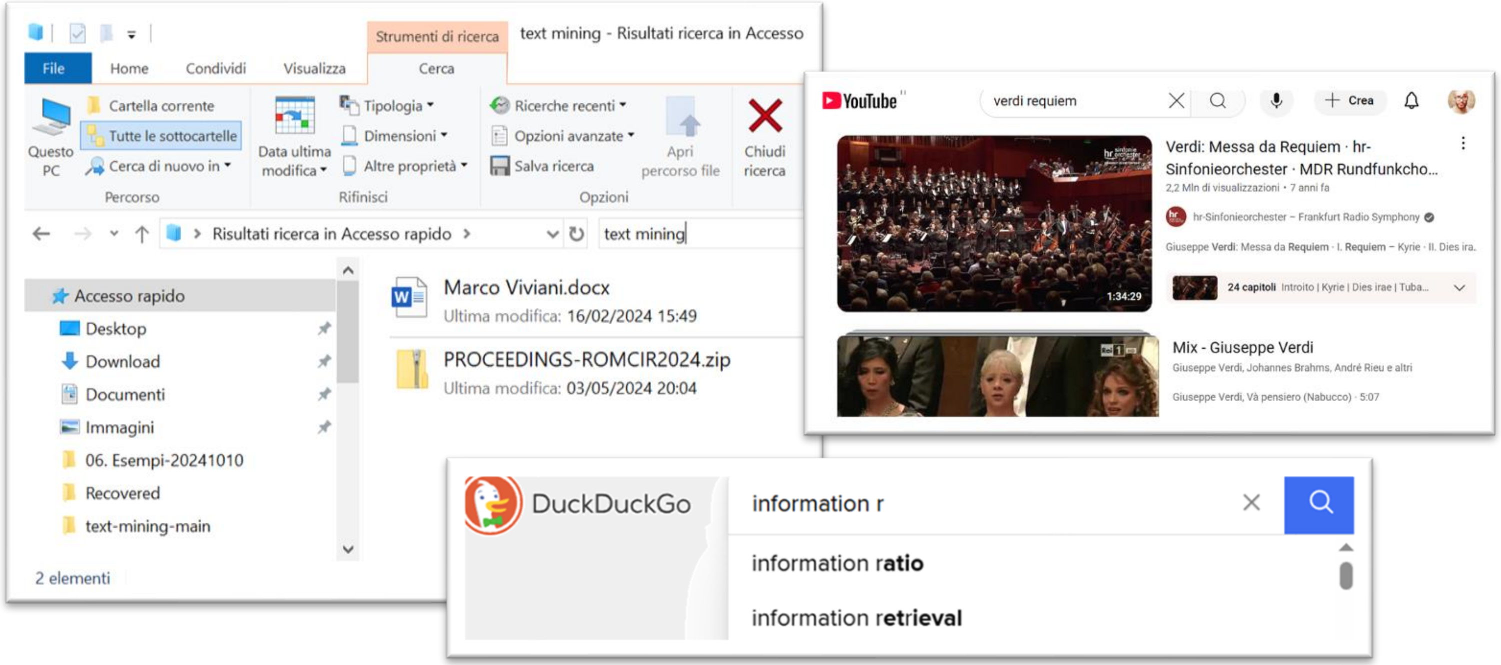
Retrieval-Augmented Text Generation

Open-book exam

Information Retrieval

- **Information Retrieval** (IR) is the process of retrieving unstructured content (typically text) from large collections to satisfy a user's information need

- **Distinct forms** of IR
  - Desktop search
  - Web search
  - Vertical search
    - Video search
    - Audio search
    - …
  - …



text mining

Marco Viviani.docx
Ultima modifica: 16/02/2024 15:49

PROCEEDINGS-ROMCIR2024.zip
Ultima modifica: 03/05/2024 20:04

verdi requiem

Verdi: Messa da Requiem · hr-Sinfonieorchester · MDR Rundfunkcho...
2,2 Mln di visualizzazioni • 7 anni fa
hr-Sinfonieorchester – Frankfurt Radio Symphony
Giuseppe Verdi: Messa da Requiem · I. Requiem – Kyrie · II. Dies ira.

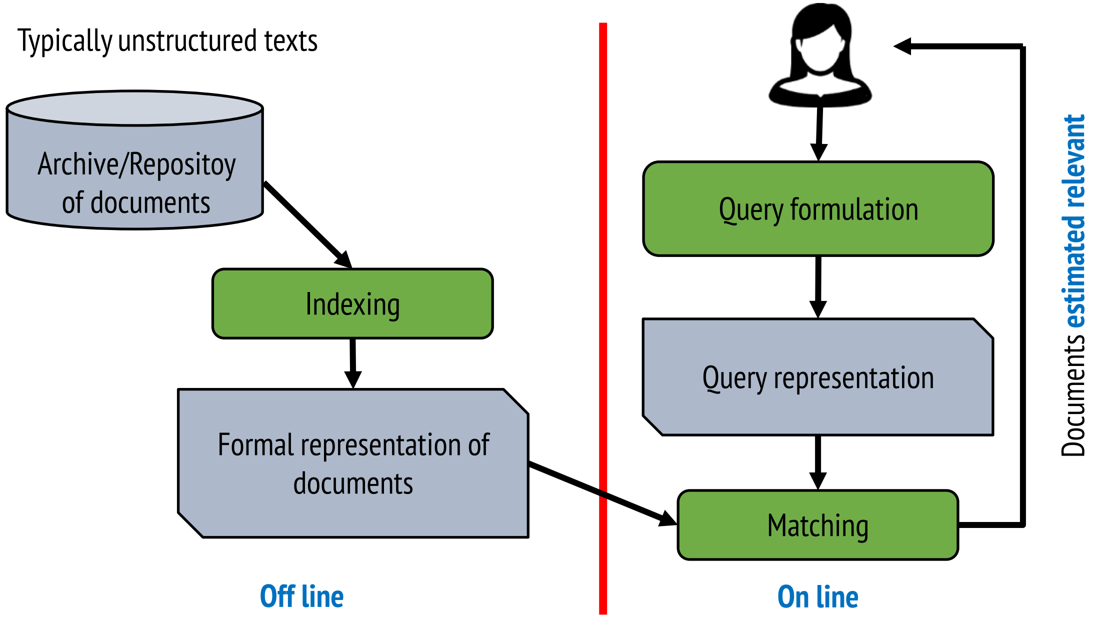
Mix - Giuseppe Verdi
Giuseppe Verdi, Johannes Brahms, André Rieu e altri
Giuseppe Verdi, Và pensiero (Nabucco) · 5:07

information r
information **ratio**
information **retrieval**

Search Engines and Relevance

- Any Search Engine is based on a **mathematical model** (IR model) that provides a formal description:
  - of the document and the query
  - of how to compare the query and the document representations to **estimate the relevance** of documents to the query

- The **relevance** of a document is relative to the formulated query (i.e., topical relevance, a.k.a. topicality)
  - Nowadays → Multi-dimensional relevance, i.e., topicality "+" novelty, popularity, factual accuracy, …


Typically unstructured texts
Archive/Repositoy of documents
Indexing
Formal representation of documents
Off line
Query formulation
Query representation
Matching
On line
Documents estimated relevant

Sparse VS Dense Retrieval

Sparse Retrieval

- Represents queries/docs as sparse vectors (mostly zeros)
  - Bag-of-Words, TF-IDF
- Based on term matching (e.g., VSM, BM25)
- Relies on exact keyword overlaps
- Fast and interpretable
- Limited in capturing semantic meaning

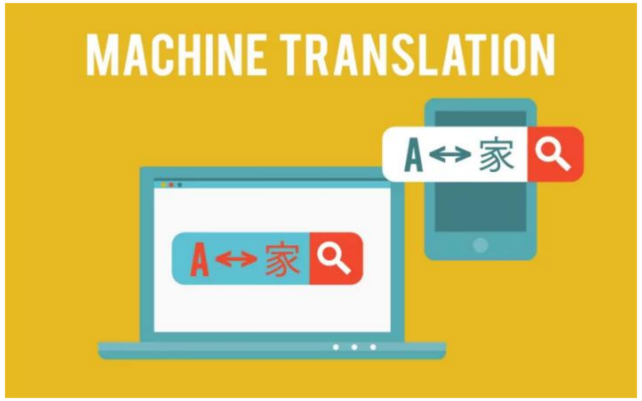
Dense Retrieval

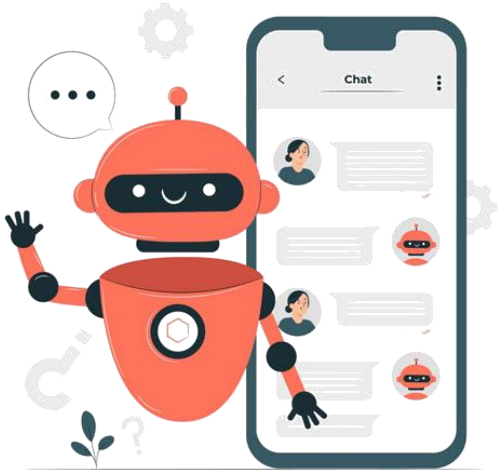
- Represents queries/docs as dense vectors
  - Neural embeddings (e.g., BERT-based)
- Captures semantic similarity, not just keywords
- Requires more compute (Approximate Nearest Neighbor search)
- Often more effective in open-domain QA and semantic search

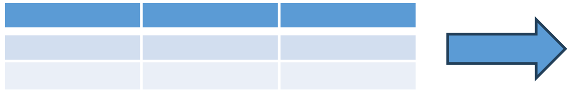
Text Generation

- **Text generation**, also known as **Natural Language Generation** (NLG), is the task of automatically producing coherent and contextually relevant text, to approximate or replicate human-written language

- Several applications
    - Machine translation
    - Open-ended text generation
    - Summarization
    - Dialogue generation / Chatbots
    - Data-to-text generation
    - …





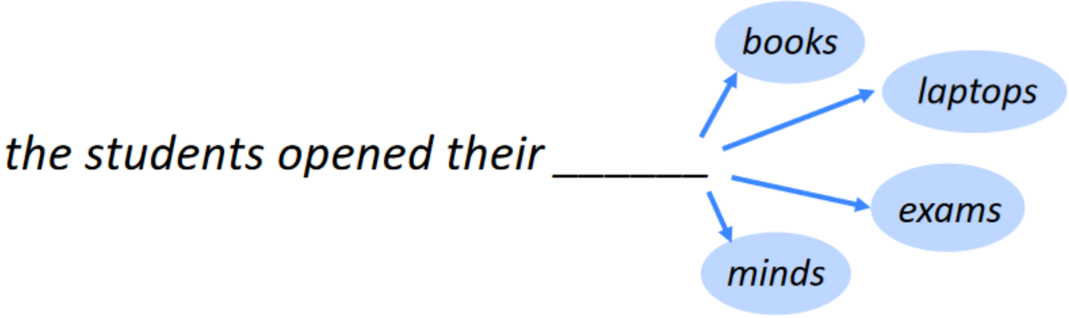


Text text text text text text text …

NLG and Language Modeling

- Natural Language Generation (NLG) and **Language Modeling** (LM) are deeply connected
  - A language model is a **probabilistic model** that estimates the likelihood of a given sequence of words

- It can be regarded as a probabilistic mechanism for **“generating” text**, thus also called a “generative” model
  - The language model learns to predict the next word given the previous ones
    - $P(w_t \mid w_1, w_2, \dots, w_{t-1})$



Before LLMs

n-gram LMs

- Collect statistics about how frequent different n-grams are (Auto-regressive LM / Causal Language Modeling)
  - Hard to compute the probability of unseen text
  - Need to store count for all n-grams
    Increasing n or corpus increases model size!
  - Language has long-distance dependencies:
    "The computer which I had just put into the machine room on the fifth floor crashed"

Neural LMs

- Use neural networks to learn word representations and model longer dependencies
  - Word2Vec / GloVe (Static embeddings)
  - RNNs / LSTMs (Model sequences better than n-grams)
  - Still struggles with long-term context, sequential computation is slow

Large Language Models (LLMs)

- **Attention-based architectures** (e.g., Transformer by Vaswani et al., 2017)
- Unlike earlier models (like RNNs or LSTMs), attention allows for **parallel computation** and **long-range dependencies** in sequences
- **Massive scale**: Billions of parameters, trained on diverse and massive corpora
    - **Knowledge** is baked into weights
    - **Self-supervised learning.** No labeled data needed

- **GPT (Generative Pre-trained Transformer)**:
    - GPT-1 (2018) → 117 million parameters, 985 million words
    - GPT-2 (2019) → 1.5 billion parameters
    - GPT-3 (2020) → 175 billion parameters. Chat GPT is also based on this model
    - GPT-4 (early 2023) → likely to contain trillions of parameters
    - GPT-4 Turbo (late 2023), optimized for efficiency → unspecified parameter count

https://www.geeksforgeeks.org/large-language-model-llm/

General-purpose, but not always task- or domain-specific

Optimizing LLMs

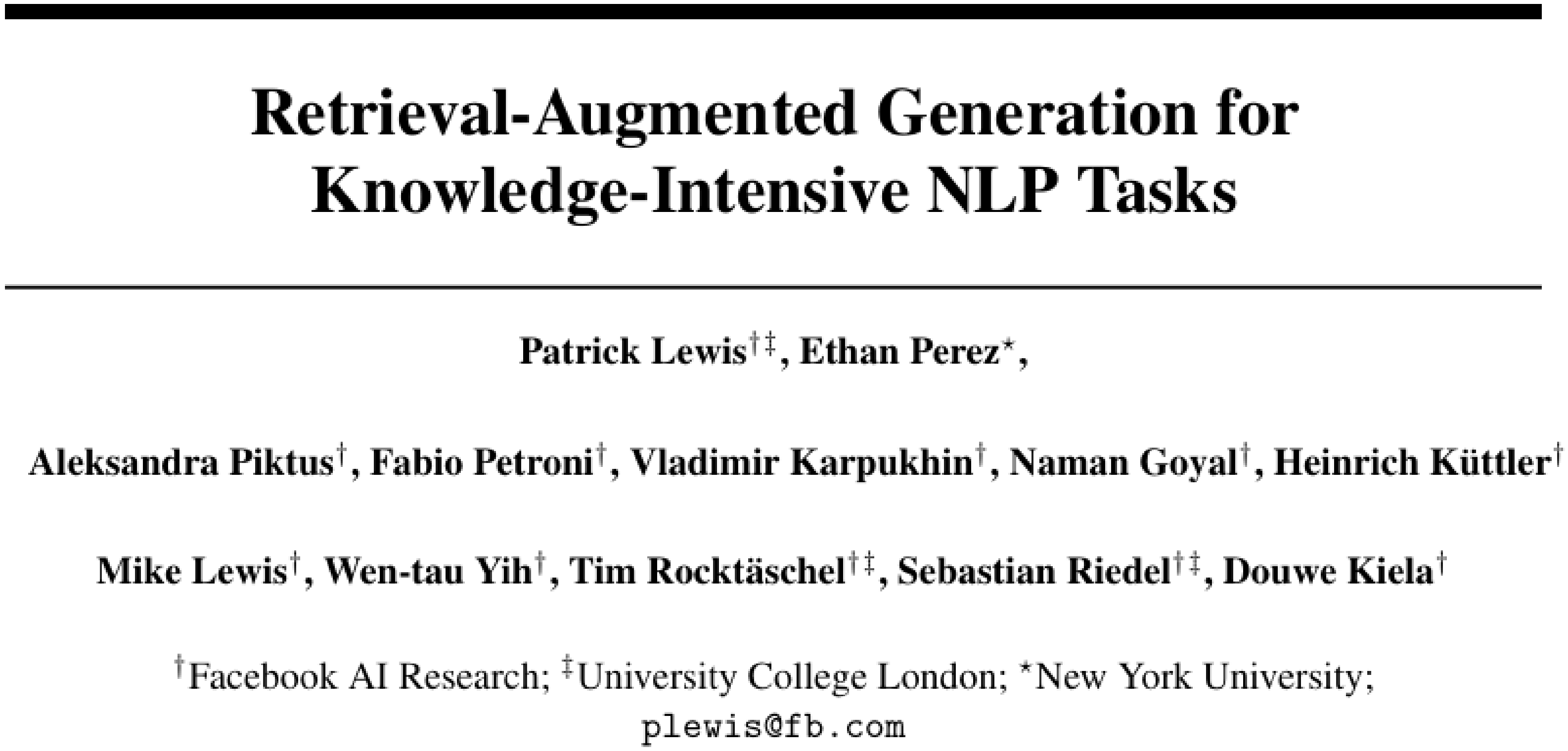
- **Fine-tuning** → Adapting a pre-trained model to a specific task or domain by training it further on a new, usually smaller, dataset
  - The model's weights are updated to perform better on the new task
    - Task-specific fine-tuning: Like text classification, question answering, or summarization
    - Domain-specific fine-tuning: Like medical, legal, or technical text
  - Costly, data-hungry, hard to update knowledge

- **In-context learning** → Teaching a language model to perform a task just by showing examples or instructions in the input **prompt**
  - No need to change model weights – just craft clever inputs (“prompts”) to guide the model
  - Zero-/Few-shot learning → Generalize with minimal examples
  - Short-term → The model "learns" the task only during the current interaction
  - Not scalable for large systems

A possible solution?

Retrieval-Augmented Generation

The Emergence of the Concept (2020)



Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

Patrick Lewis[†‡], Ethan Perez[*],

Aleksandra Piktus[†], Fabio Petroni[†], Vladimir Karpukhin[†], Naman Goyal[†], Heinrich Küttler[†]

Mike Lewis[†], Wen-tau Yih[†], Tim Rocktäschel[†‡], Sebastian Riedel[†‡], Douwe Kiela[†]

[†]Facebook AI Research; [‡]University College London; [*]New York University;
plewis@fb.com

https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

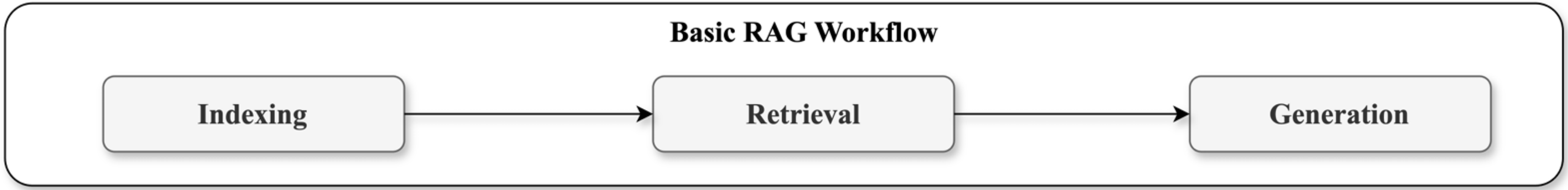
RAG: Basic Notions

- The idea behind RAG techniques is to **make use of knowledge "outside" the model** to provide a **"local" context (in-context)** that can supplement the model with appropriate knowledge without changing its parameters

- These are basically prompting techniques that supplement the user's input with **contextual knowledge retrieved** by accessing external sources of information through a search engine

Naive RAG: Pipeline

- During the nascent stages of RAG, its core framework is constituted by **indexing**, **retrieval**, and **generation**, a paradigm referred to as **Naive RAG**



Basic RAG Workflow

Indexing → Retrieval → Generation

Naive RAG: Indexing and Retrieval

- **Indexing**: creating an inverted index–mapping each token to the documents/positions where it appears
  - We often tokenize and chunk documents → Each chunk is a set of tokens that can fit within the model’s context window
  - We generate embeddings for those chunks and index those

- **Retrieval**
  - BM25 and other sparse IR models focus on term frequency and presence for document ranking → they often overlook the semantic information of queries
  - Current strategies leverage dense IR models based on pretrained LMs like BERT → they capture the semantic essence of queries more effectively

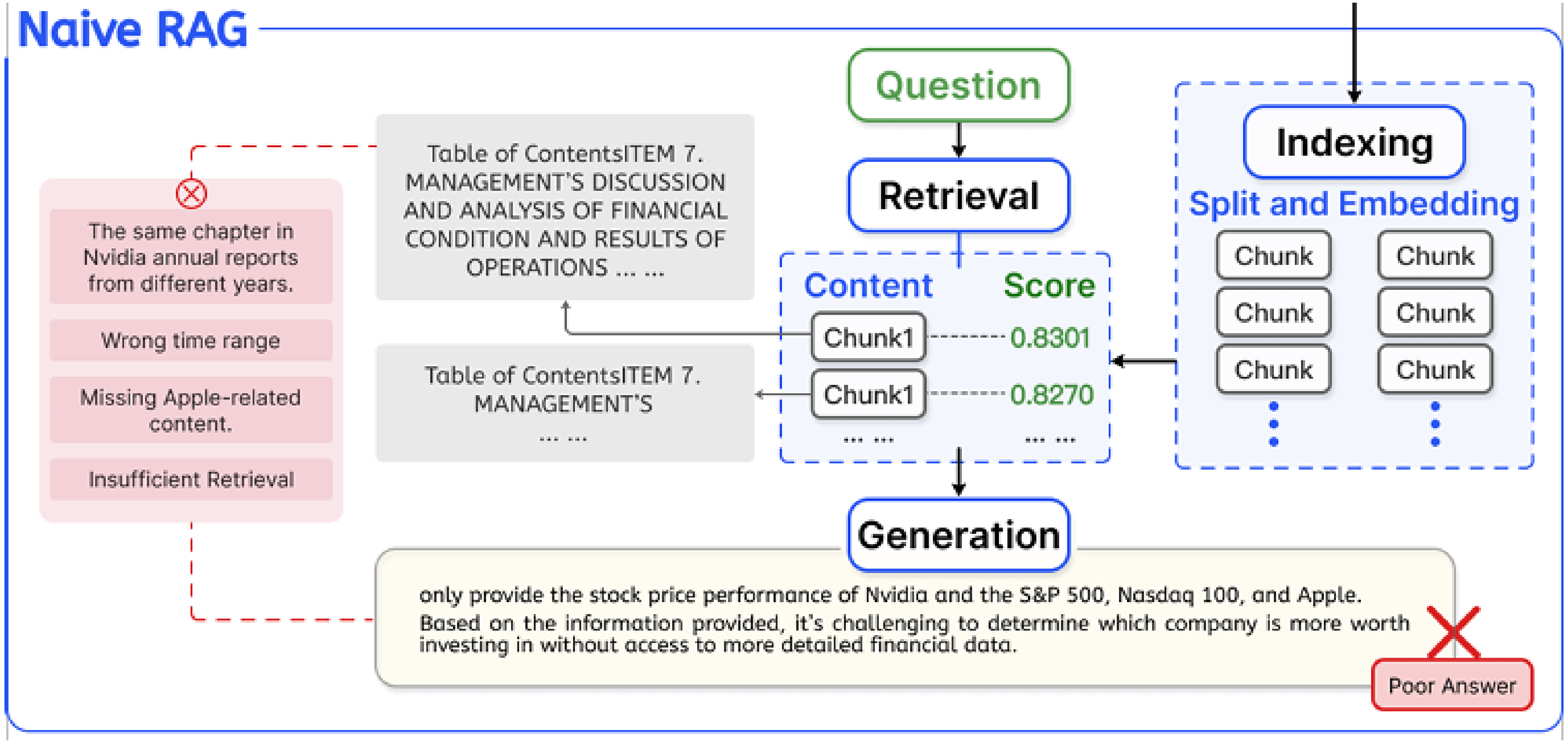
Naive RAG: Generation

- The generation phase is tasked with producing text that is both **relevant to the query** and **reflective of the information** found in the retrieved documents

- The usual method involves **concatenating the query with the retrieved information**, which is then **fed into an LLM for text generation**

- The generated text should accurately convey the information from the retrieved documents and align with the query's intent, while also offering the flexibility to introduce new insights or perspectives not explicitly contained within the retrieved data

Naive RAG: An Example

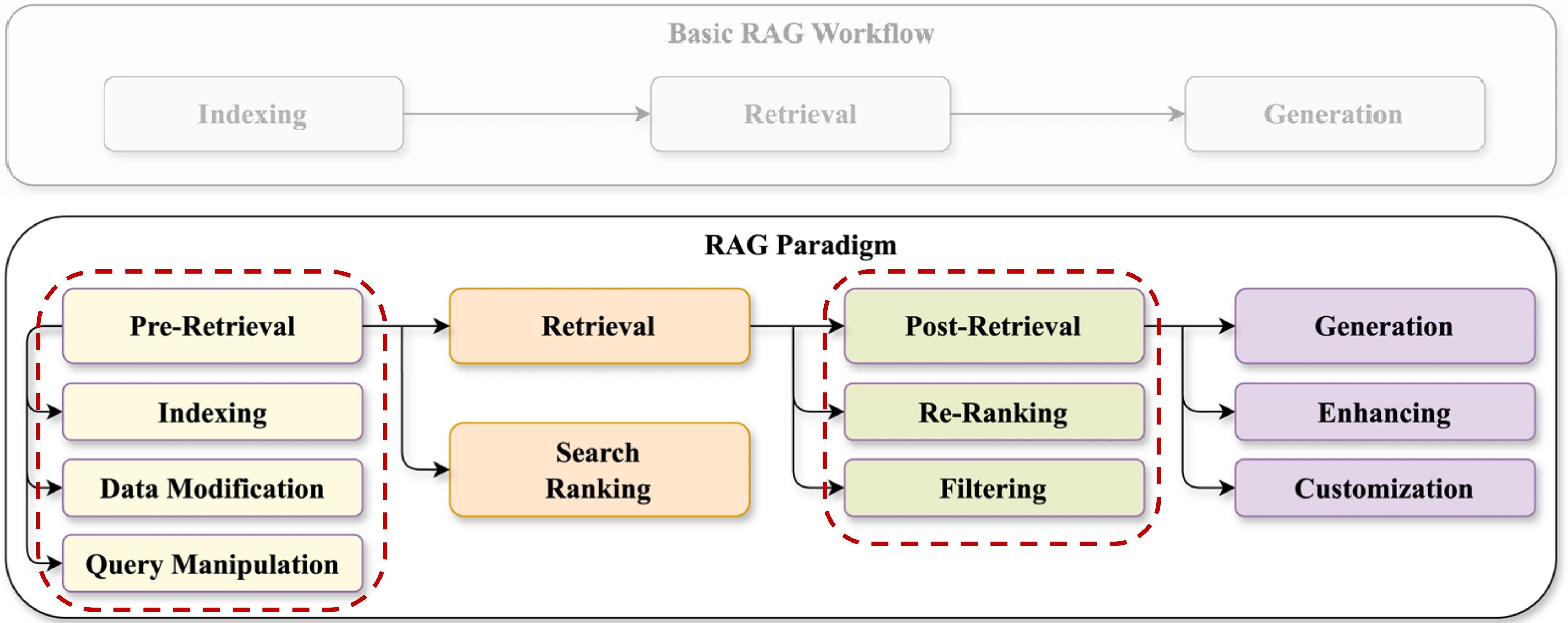

Naive RAG
Question
Retrieval
Indexing
Split and Embedding
Chunk
Chunk
Chunk
Chunk
Chunk
Chunk
Content
Score
Chunk1
0.8301
Chunk1
0.8270
... ...
... ...
Table of ContentsITEM 7. MANAGEMENT'S DISCUSSION AND ANALYSIS OF FINANCIAL CONDITION AND RESULTS OF OPERATIONS ... ...
Table of ContentsITEM 7. MANAGEMENT'S ... ...
The same chapter in Nvidia annual reports from different years.
Wrong time range
Missing Apple-related content.
Insufficient Retrieval
Generation
only provide the stock price performance of Nvidia and the S&P 500, Nasdaq 100, and Apple. Based on the information provided, it's challenging to determine which company is more worth investing in without access to more detailed financial data.
Poor Answer

Advanced RAG

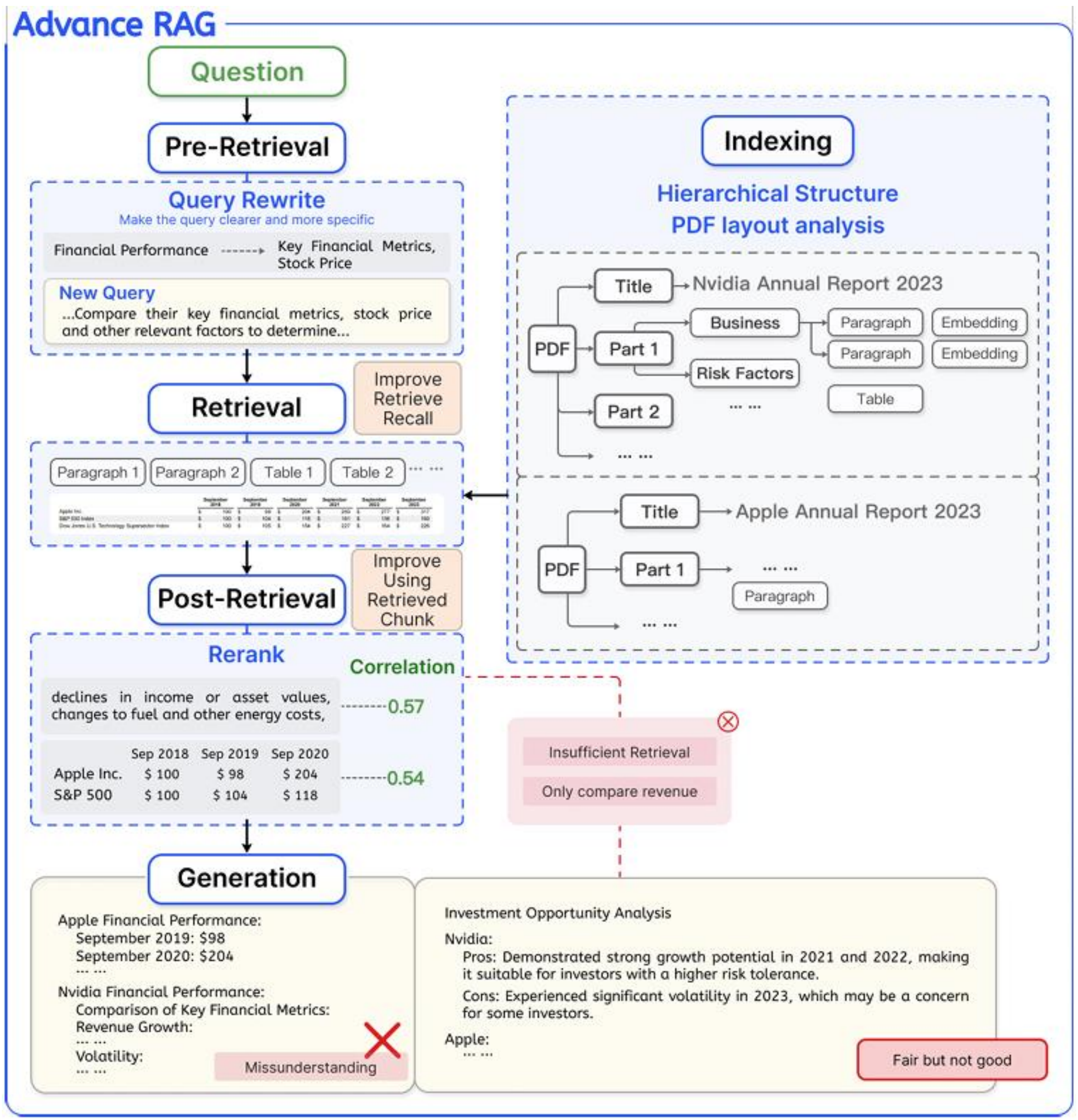
- Advanced RAG focuses on **optimizing the retrieval phase**, aiming to enhance retrieval efficiency and strengthen the utilization of retrieved chunks
  - Research indicates that an excess of redundant and noisy information may interfere with the LLM's identification of key information

- Typical strategies involve **pre-retrieval processing** and **post-retrieval processing**
  - The specificity of indexing depends on the task and data type (e.g., sentence-level indexing or paragraph-level indexing is better for Q-A systems)
  - Avoiding data redundance
  - Query rewriting is used to make the queries clearer and more specific, thereby increasing the accuracy of retrieval
  - The reranking/filtering of retrieval results is employed to enhance the LLM's ability to identify and utilize key information

Advanced RAG: Pipeline



Advanced RAG: An Example



Financial Performance -------→ Key Financial Metrics, Stock Price

New Query
...Compare their key financial metrics, stock price and other relevant factors to determine...

declines in income or asset values, changes to fuel and other energy costs, --------0.57

	Sep 2018	Sep 2019	Sep 2020
Apple Inc.	$ 100	$ 98	$ 204
S&P 500	$ 100	$ 104	$ 118

--------0.54

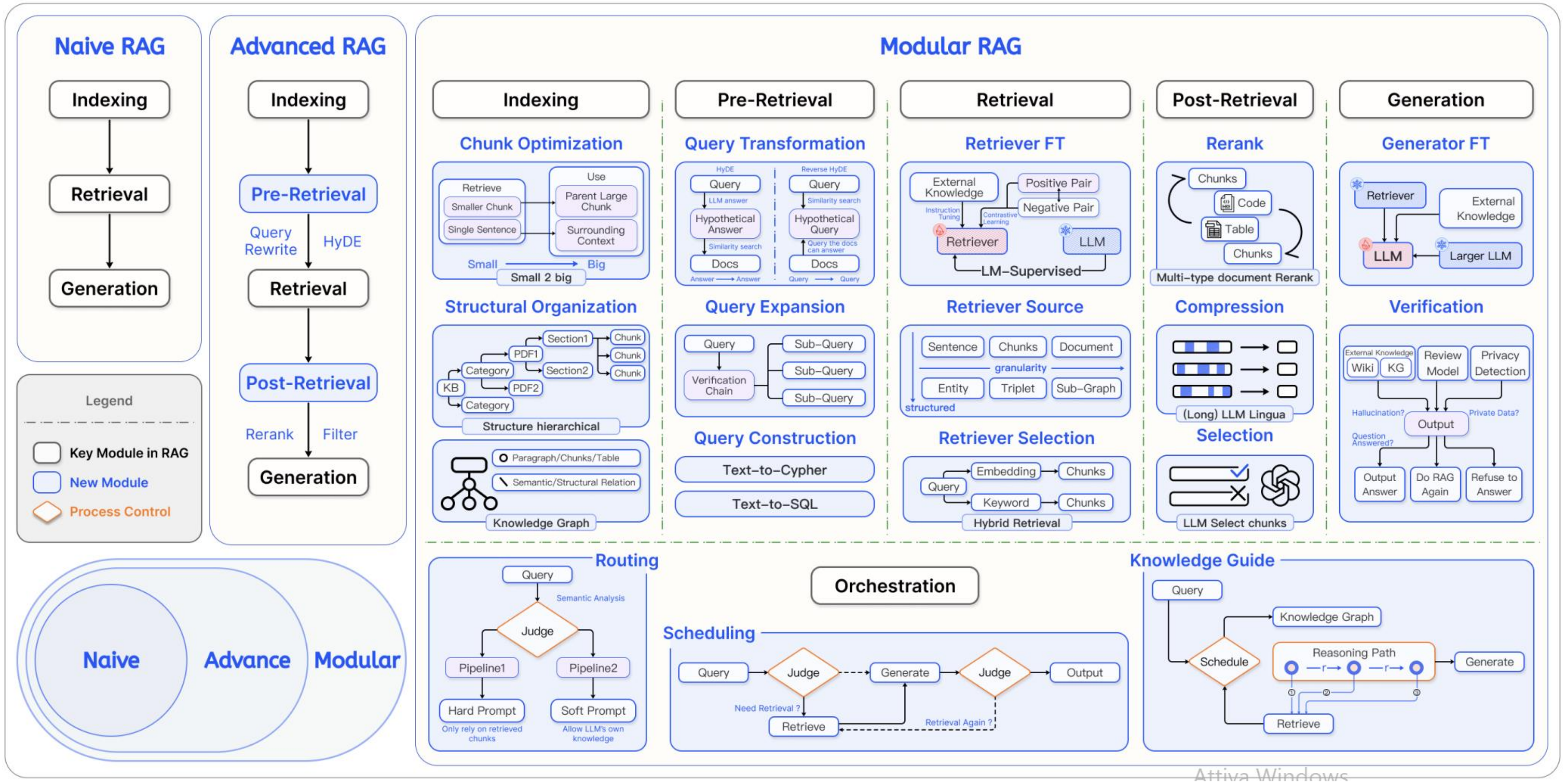
Apple Financial Performance:
September 2019: $98
September 2020: $204
··· ···
Nvidia Financial Performance:
Comparison of Key Financial Metrics:
Revenue Growth:
··· ···
Volatility:
··· ···

Investment Opportunity Analysis
Nvidia:
Pros: Demonstrated strong growth potential in 2021 and 2022, making it suitable for investors with a higher risk tolerance.
Cons: Experienced significant volatility in 2023, which may be a concern for some investors.
Apple:
··· ···

Modular RAG

- The **current RAG paradigm** → Surpassing the traditional linear retrieval-generation paradigm
- **Modular RAG** → Consists of multiple independent yet tightly coordinated modules, each responsible for handling specific functions or tasks
- **Advantages of Modular RAG** → It enhances the flexibility and scalability of RAG systems
    - Users can flexibly combine different modules and operators according to the requirements of data sources and task scenarios

Modular RAG: Pipeline



Modular RAG: Orchestration

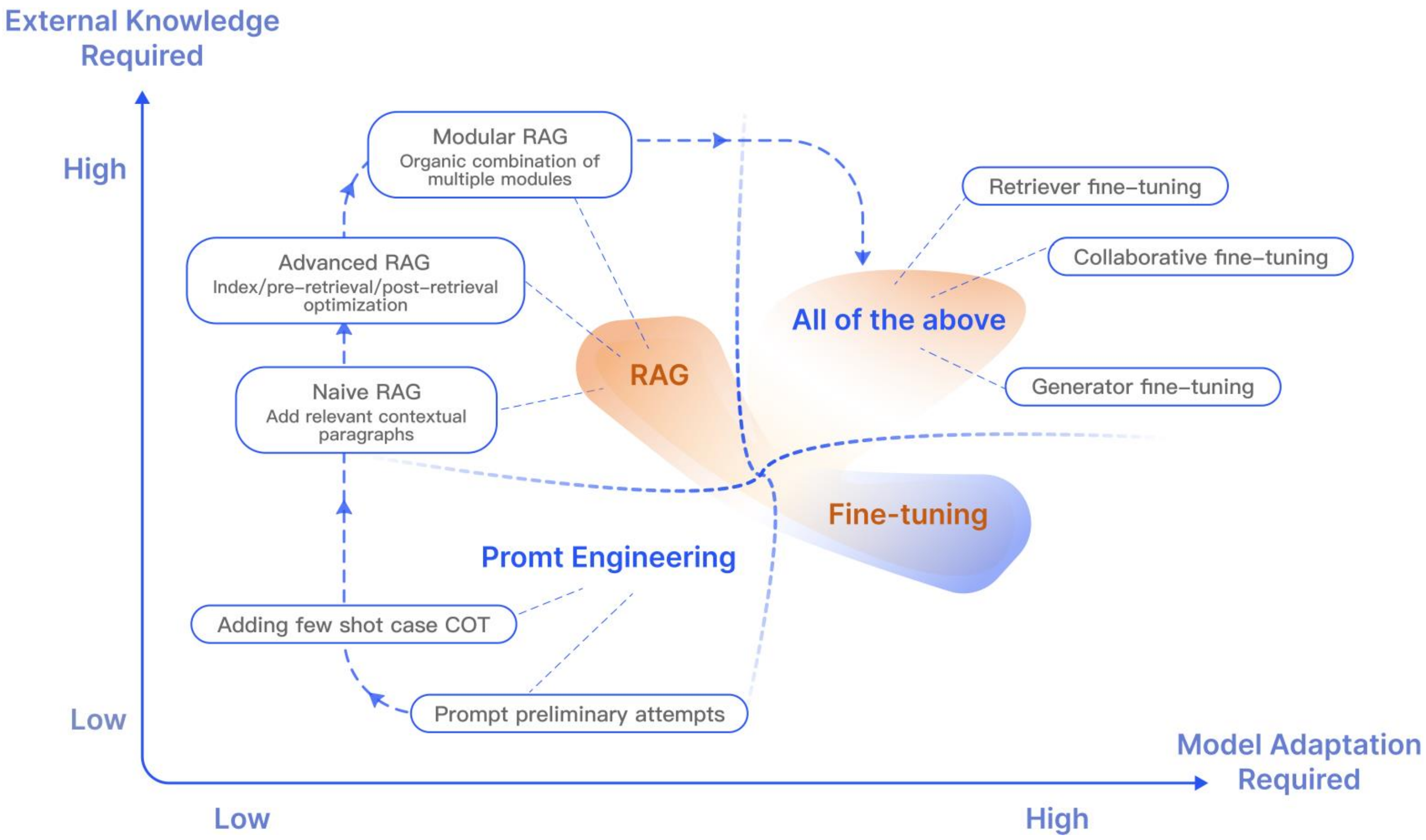
- Modular RAG incorporates **decision-making at pivotal junctures** and dynamically selects subsequent steps contingent upon the previous outcomes

- Routing
  - In response to diverse queries, the RAG system routes to specific pipelines tailored for different scenario, a feature essential for a versatile RAG architecture designed to handle a wide array of situations

- Fusion
  - Enhancing diversity by exploring multiple pipelines → Fusing for the best output

- Scheduling
  - It identifies critical junctures that require external data retrieval, assessing the adequacy of the responses, and deciding on the necessity for further investigation
  - It is commonly utilized in scenarios that involve recursive, iterative, and adaptive retrieval

RAG vs ALL


External Knowledge
Required
High
Low
Modular RAG
Organic combination of multiple modules
Advanced RAG
Index/pre–retrieval/post–retrieval optimization
Naive RAG
Add relevant contextual paragraphs
RAG
Retriever fine–tuning
Collaborative fine–tuning
All of the above
Generator fine–tuning
Fine-tuning
Promt Engineering
Adding few shot case COT
Prompt preliminary attempts
Model Adaptation
Required
Low
High

RAG in Practice

Challenges

- Users may face **various risks** in releasing and accessing content (structured, semi-structured, unstructured) in online environments.

- **Content release**: Uncontrolled release of personal/sensitive data (privacy)
  - How to protect privacy?
  - How to avoid microtargeting?

- **Content access**: Access to "incomplete"/fake information
  - How to avoid misinformation access?

- Provide users with automated and effective approaches promoting user autonomy with easily interpretable results without the decision-making process being left only to algorithms

→ **KURAMi**

The KURAMi Project

- **KURAMi**: Knowledge-based, explainable User empowerment in Releasing private data and Assessing Misinformation in online environments

- **PRIN 2022**: Research project funded by the Italian EU - Next Generation EU, Mission 4, Component 2, CUP D53D23008480001 and Italian MUR

Finanziato dall'Unione europea
NextGenerationEU

Ministero dell'Università e della Ricerca

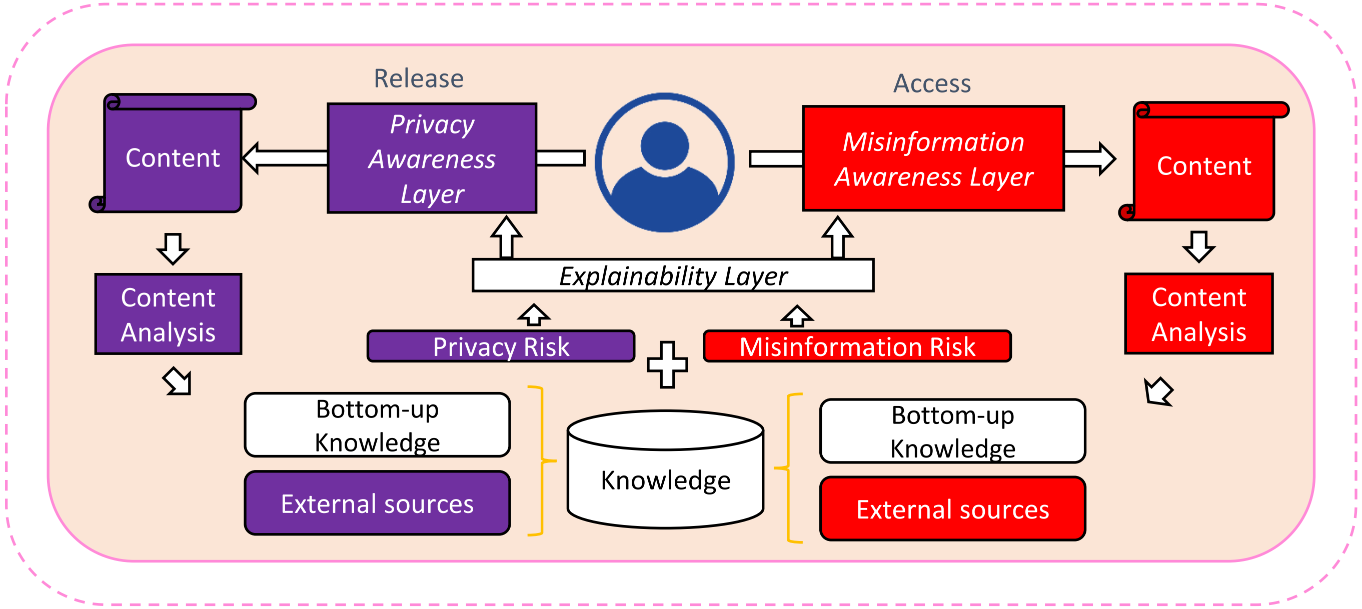
Italiadomani
PIANO NAZIONALE DI RIPRESA E RESILIENZA

KURAMi: Some Tasks

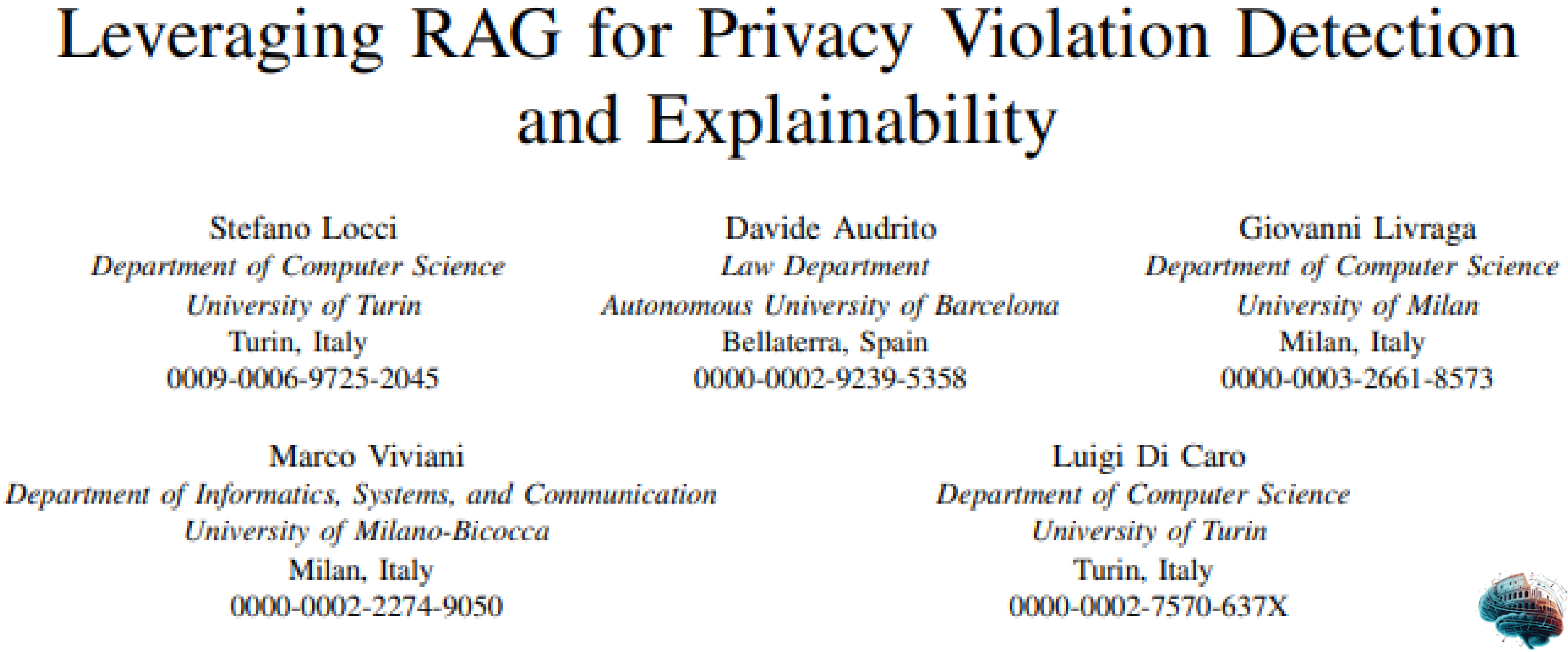
- Various **tasks** are involved in KURAMi

- In today's **seminar**:
  - Privacy Awareness
    - RAG-based privacy violation detection
  - Misinformation Awareness
    - RAG-based «truthful» health IR

KURAMi-ENRICHED ONLINE ENVIRONMENT



RAG for Privacy Violation Detection



Leveraging RAG for Privacy Violation Detection and Explainability

Stefano Locci
Department of Computer Science
University of Turin
Turin, Italy
0009-0006-9725-2045

Davide Audrito
Law Department
Autonomous University of Barcelona
Bellaterra, Spain
0000-0002-9239-5358

Giovanni Livraga
Department of Computer Science
University of Milan
Milan, Italy
0000-0003-2661-8573

Marco Viviani
Department of Informatics, Systems, and Communication
University of Milano-Bicocca
Milan, Italy
0000-0002-2274-9050

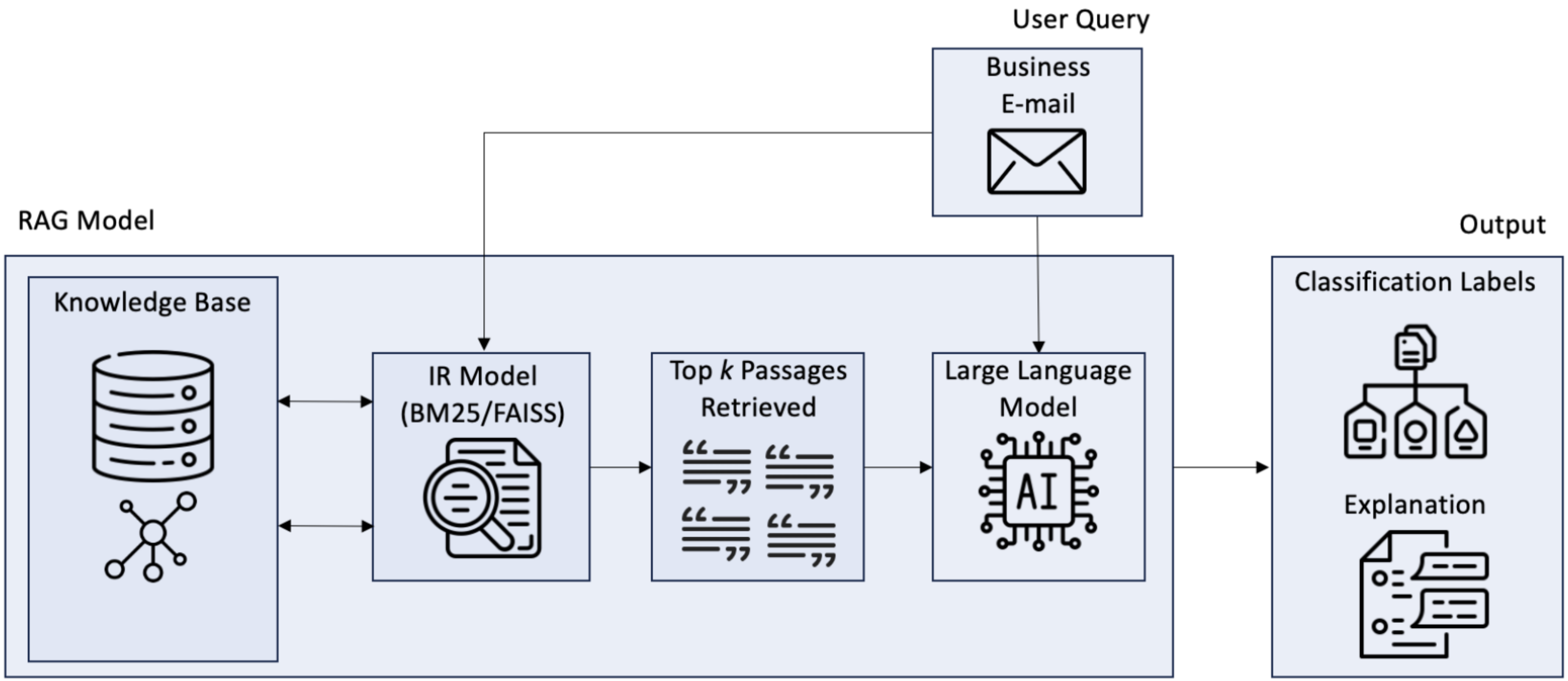
Luigi Di Caro
Department of Computer Science
University of Turin
Turin, Italy
0000-0002-7570-637X

INTERNATIONAL JOINT CONFERENCE ON NEURAL NETWORKS
IJCNN2025
30 JUNE - 5 JULY 2025 | ROME, ITALY
International Neural Network Society

The Proposed Solution

- A **RAG-based framework for privacy violation detection** with clear natural language explanations
- A curated domain-specific **Knowledge Base** (KB)
- **Annotation** and **Evaluation** on the Enron Email Dataset

The Proposed Solution: Pipeline



The Curated Knowledge Base

- **100 Scientific Privacy-Related Papers**
  - Sources: PETS, IEEE S&P, ACM CCS, WPES
  - Focused on privacy-sensitive data, confidentiality, and data protection in the context of AI/ML
- **Filtered for Technical Insights**
  - Articles addressing privacy risks, sensitive data identification, privacy-preserving technologies
  - Excluding full GDPR guidelines or general legal frameworks
- **Three KB versions**
  - Only Abstracts
  - Only Introductions
  - Introductions + Abstracts

The Classification Task

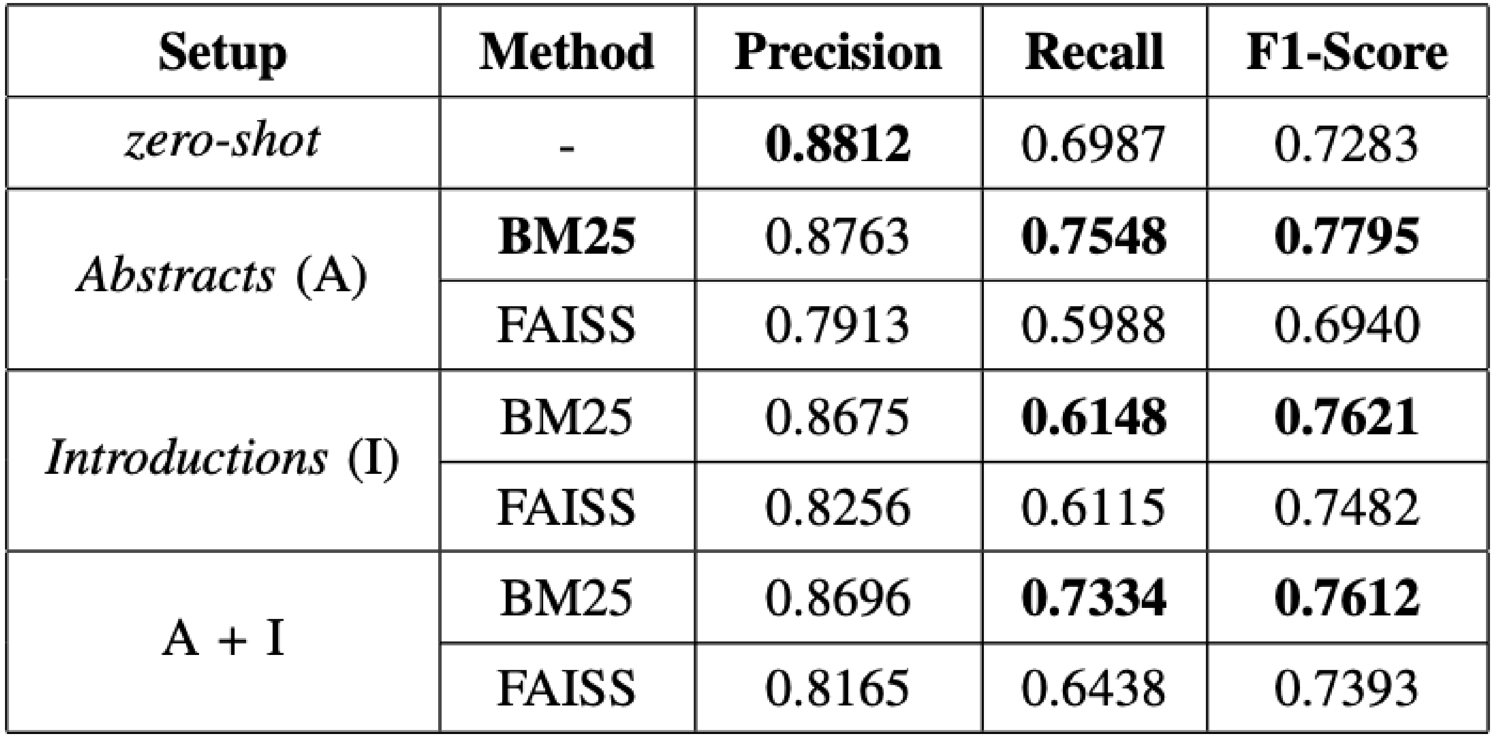
- **Task**: classify emails for privacy violations by identifying potentially **sensitive information**

- Privacy-sensitive information includes:
  - **Personally Identifiable Information (PII)**: Information that can directly identify an individual (e.g., names, addresses, social security numbers)
  - **Confidential business data**: Confidential data that belongs to a company, such as internal emails, business strategies, or proprietary data
  - **Credentials**: Sensitive data like usernames, passwords, or access codes

Dataset Annotation

- The **Enron Email Dataset**
  - https://www.kaggle.com/datasets/wcukierski/enron-email-dataset
- **120 emails** annotated by **human experts** and **GPT-4o** (few-shot learning approach)
- **4 human-annotated examples** provided to GPT-4o for few-shot learning.
  - Cohen's kappa score: 0.8 (substantial agreement)
- **1,103 emails** labeled automatically by **GPT-4o**

Some Results



Setup	Method	Precision	Recall	F1-Score
zero-shot	-	**0.8812**	0.6987	0.7283
Abstracts (A)	**BM25**	0.8763	**0.7548**	**0.7795**
	FAISS	0.7913	0.5988	0.6940
Introductions (I)	BM25	0.8675	**0.6148**	**0.7621**
	FAISS	0.8256	0.6115	0.7482
A + I	BM25	0.8696	**0.7334**	**0.7612**
	FAISS	0.8165	0.6438	0.7393

A Tool for Explainability? (Zero-Shot)

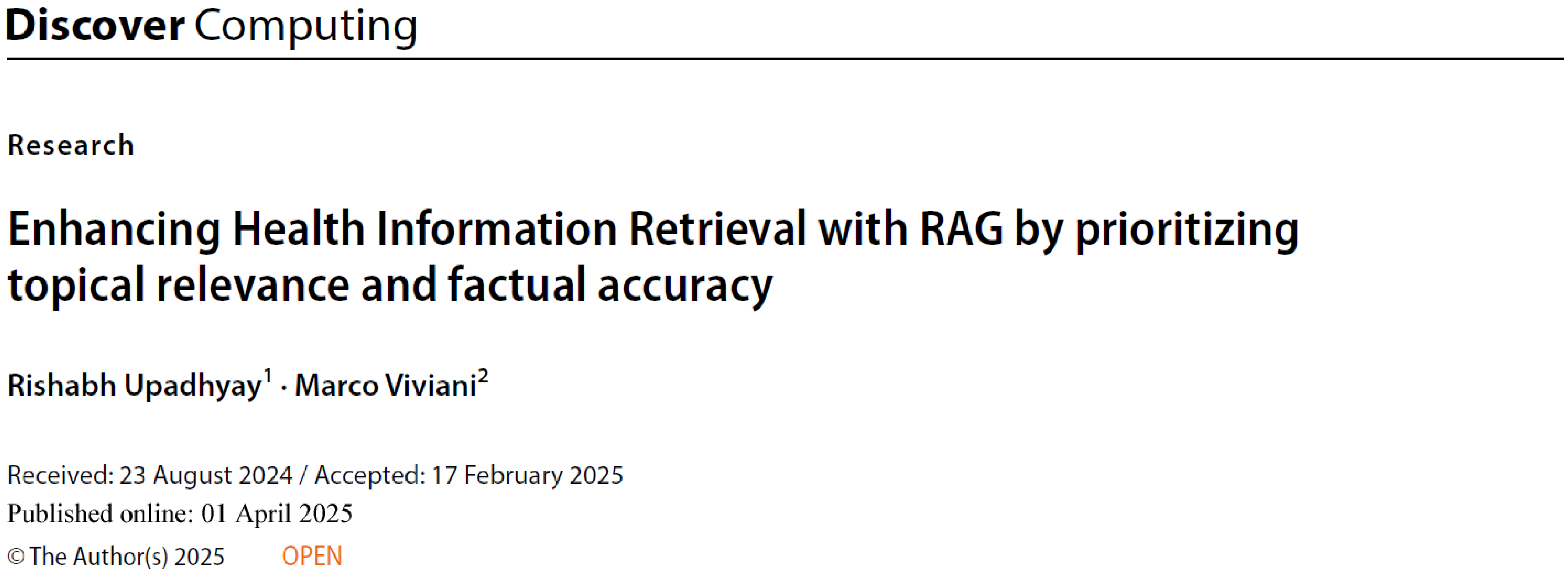
- **Email**: "…availability of your current Natural Gas Intelligence Newsletter(s). Please use your username of 'kward' and your password to access…"
- **Label**: Violation detected
- **Explanation**: The text contains sensitive information, specifically the username 'kward' and a password for accessing subscription-based services provided by Natural Gas Intelligence. Sharing login credentials publicly or through unsecured channels can lead to unauthorized access to personal accounts and services, which is a violation of privacy and security protocols.

A Tool for Explainability? (BM25/FAISS)

- **Email**: "…Could you please give me a Yes, No, Maybe decision on the following candidates interviewed: Bruce Kamich, Philip Roan, Jerzy Jarosz, Bruce James, Oleg Bondar. Thanks, Toni…"
- **Label**: Violation detected
- **Explanation**: According to privacy policies and data protection regulations, such as the **General Data Protection Regulation (GDPR)** in the EU or the **California Consumer Privacy Act (CCPA)** in the US, the text contains sensitive information related to the recruitment process at Enron North America Corp. It includes the names of several individuals who were candidates for a position, as well as internal communications discussing their evaluations. This information can be considered **Personally Identifiable Information (PII)** because it identifies specific individuals in the context of a job application process.

- **Email**: "Thursday's (4/6/00) conference call regarding the weekly Japan update will now be held at 6:00 p.m. Houston time and 8:00 a.m. Tokyo time due to Daylight Savings time change. Please forward your updates to Jope Hirl and copy me. The call will be held in Conference room 21C2 for Houston participants. Thanks, Eddie Robles Domestic: 1-800-991-9019 Intern'l 847-619-8039 Code: 5764488"
- **Label**: Violation detected
- **Explanation**: The text provided contains sensitive information that could potentially violate privacy policies. The following points highlight why the text is considered sensitive:
  - **Personal Identifiable Information (PII)**: The mention of the conference call details, including participants' names, could reveal personal identifiers.
  - **Contact Information**: The inclusion of phone numbers (domestic and international) adds an extra layer of privacy sensitivity.
  - **Conference Call Details**: Information about the conference call, including access codes and other logistical data, can be considered sensitive.
  - **Time and Location Information**: The timing and location of the meeting are also private data that could lead to security risks or breaches.

RAG for “Truthful” Health Information Retrieval



Discover Computing

Research

Enhancing Health Information Retrieval with RAG by prioritizing topical relevance and factual accuracy

Rishabh Upadhyay[1] · Marco Viviani[2]

Received: 23 August 2024 / Accepted: 17 February 2025
Published online: 01 April 2025

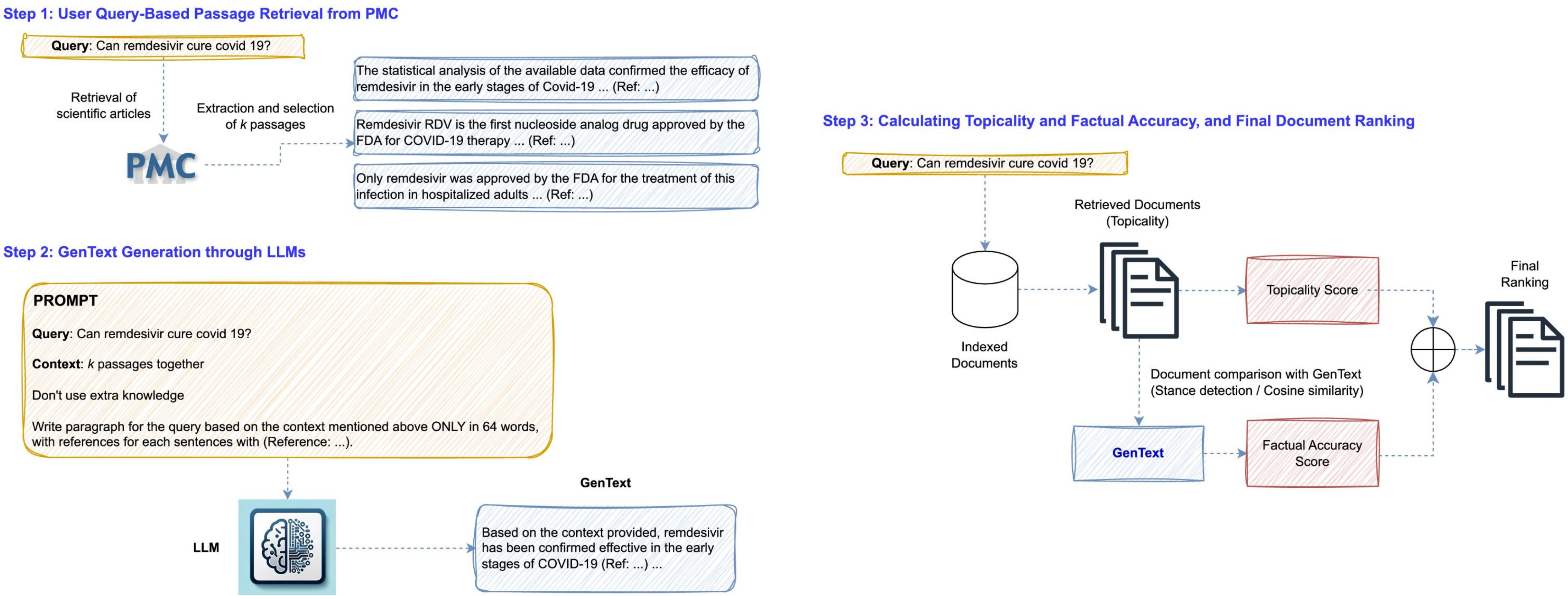
https://doi.org/10.1007/s10791-025-09505-5

The Proposed Solution

- **Integrating generative LLMs with a reputed, external knowledge**, such as the curated scientific repository of PubMed Central (PMC), a strategy designed to increase both the topical relevance and **factual accuracy** of the retrieved documents

- The proposed solution is characterized by **three key stages**:
  - User query-based passage retrieval from PMC
  - GenText generation through LLMs
  - Calculating topicality and factual accuracy, and final document ranking

The Proposed Solution: Pipeline



Step 1: User Query-Based Passage Retrieval from PMC

Query: Can remdesivir cure covid 19?

Retrieval of scientific articles

Extraction and selection of *k* passages

The statistical analysis of the available data confirmed the efficacy of remdesivir in the early stages of Covid-19 ... (Ref: ...)

Remdesivir RDV is the first nucleoside analog drug approved by the FDA for COVID-19 therapy ... (Ref: ...)

Only remdesivir was approved by the FDA for the treatment of this infection in hospitalized adults ... (Ref: ...)

Step 2: GenText Generation through LLMs

PROMPT

Query: Can remdesivir cure covid 19?

Context: *k* passages together

Don't use extra knowledge

Write paragraph for the query based on the context mentioned above ONLY in 64 words, with references for each sentences with (Reference: ...).

GenText

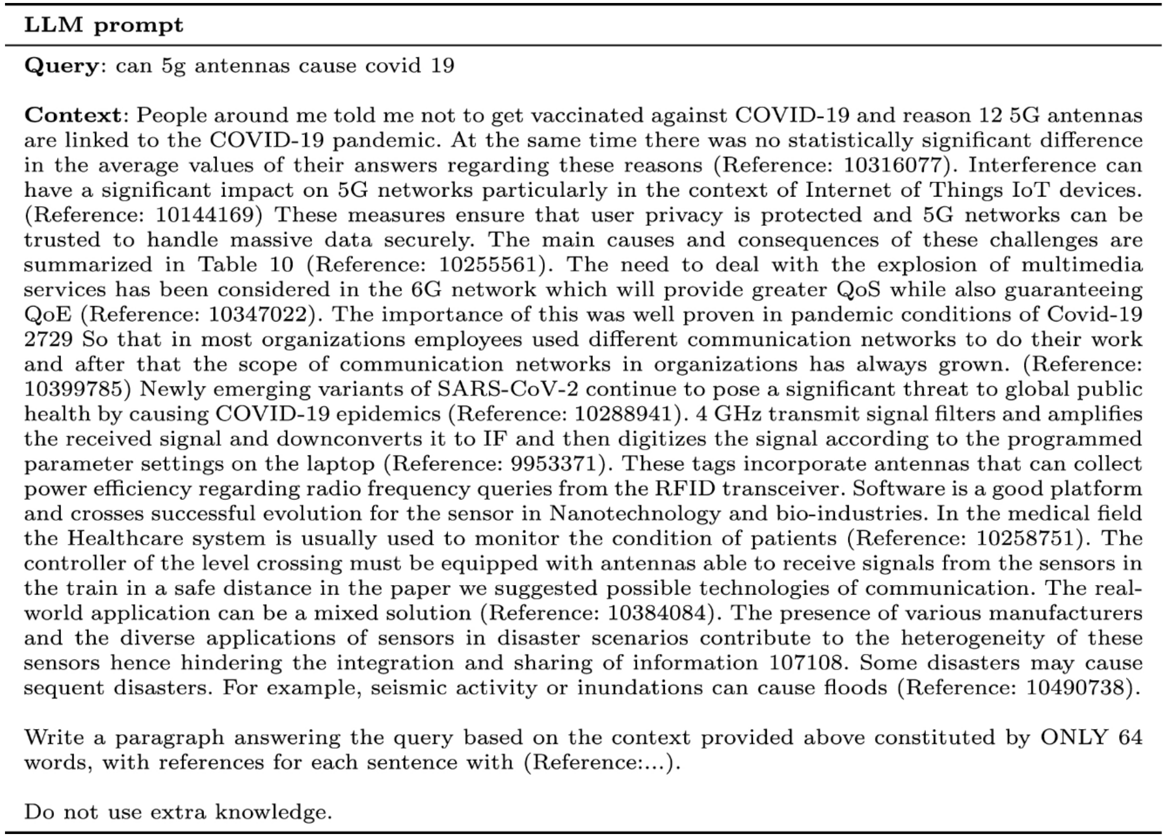
Based on the context provided, remdesivir has been confirmed effective in the early stages of COVID-19 (Ref: ...) ...

Step 3: Calculating Topicality and Factual Accuracy, and Final Document Ranking

Query: Can remdesivir cure covid 19?

Document comparison with GenText (Stance detection / Cosine similarity)

GenText generation through LLMs



LLM prompt

Query: can 5g antennas cause covid 19

Context: People around me told me not to get vaccinated against COVID-19 and reason 12 5G antennas are linked to the COVID-19 pandemic. At the same time there was no statistically significant difference in the average values of their answers regarding these reasons (Reference: 10316077). Interference can have a significant impact on 5G networks particularly in the context of Internet of Things IoT devices. (Reference: 10144169) These measures ensure that user privacy is protected and 5G networks can be trusted to handle massive data securely. The main causes and consequences of these challenges are summarized in Table 10 (Reference: 10255561). The need to deal with the explosion of multimedia services has been considered in the 6G network which will provide greater QoS while also guaranteeing QoE (Reference: 10347022). The importance of this was well proven in pandemic conditions of Covid-19 2729 So that in most organizations employees used different communication networks to do their work and after that the scope of communication networks in organizations has always grown. (Reference: 10399785) Newly emerging variants of SARS-CoV-2 continue to pose a significant threat to global public health by causing COVID-19 epidemics (Reference: 10288941). 4 GHz transmit signal filters and amplifies the received signal and downconverts it to IF and then digitizes the signal according to the programmed parameter settings on the laptop (Reference: 9953371). These tags incorporate antennas that can collect power efficiency regarding radio frequency queries from the RFID transceiver. Software is a good platform and crosses successful evolution for the sensor in Nanotechnology and bio-industries. In the medical field the Healthcare system is usually used to monitor the condition of patients (Reference: 10258751). The controller of the level crossing must be equipped with antennas able to receive signals from the sensors in the train in a safe distance in the paper we suggested possible technologies of communication. The real-world application can be a mixed solution (Reference: 10384084). The presence of various manufacturers and the diverse applications of sensors in disaster scenarios contribute to the heterogeneity of these sensors hence hindering the integration and sharing of information 107108. Some disasters may cause sequent disasters. For example, seismic activity or inundations can cause floods (Reference: 10490738).

Write a paragraph answering the query based on the context provided above constituted by ONLY 64 words, with references for each sentence with (Reference:...).

Do not use extra knowledge.

Based on the context provided, there is a misconception linking 5G antennas to the COVID-19 pandemic (Reference: 10316077). However, this connection has no statistically significant evidence to support it (Reference: 10316077). Instead, it's important to note that 5G networks play a crucial role in ensuring secure data handling and enhancing user privacy (Reference: 10255561). Moreover, SARS-CoV-2 variants remain the main cause of COVID-19 outbreaks (Reference: 10288941).

Topicality, Factual Accuracy, and Final Ranking

- The **topicality score** T is computed for each document d in the document collection, and it is derived using the BM25 retrieval model against the considered query q

$$T(d,q) = \mathrm{BM25}(d,q)$$

- The **factual accuracy score** F integrates two distinct metrics to assess the adherence of documents to be retrieved to scientific evidence within GenText G

$$F(d,G) \\ = \alpha \cdot stance(d,G) + (1-\alpha) \cdot \cos(d,G)$$

- The **final document ranking** is obtained by performing a linear combination of topicality and factual accuracy scores in order to obtain the Retrieval Status Value (RSV)

$$RSV(d,q,G) = \beta \cdot T(d,q) + (1-\beta) \cdot F(d,g)$$

Some Results

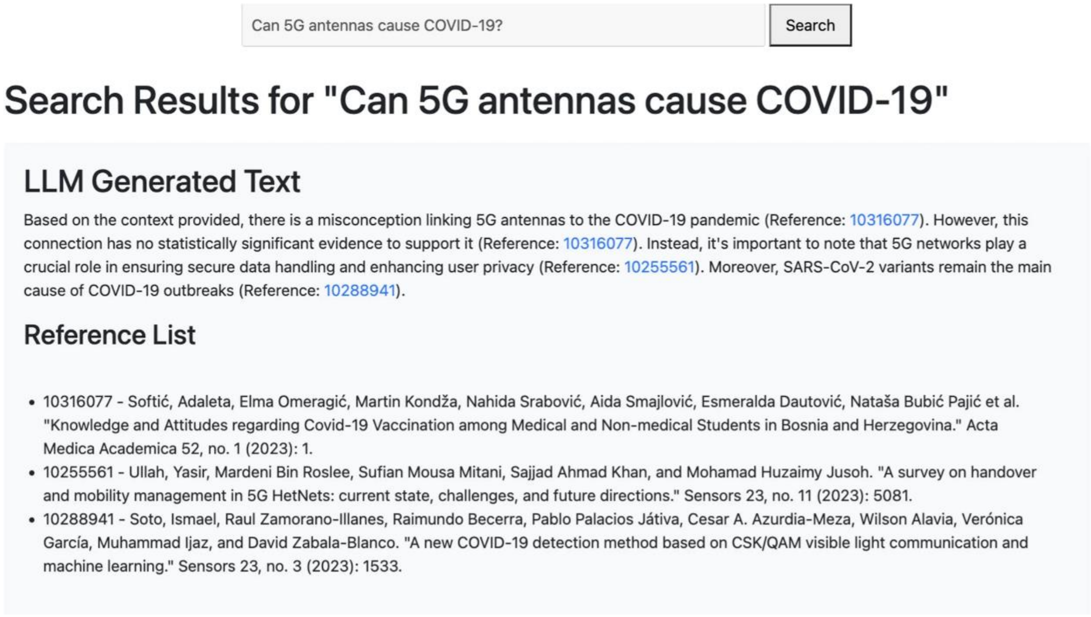
CLEF eHealth 2020 dataset

Model	CAM_{MAP}	CAM_{NDCG}	Embeddings
Top-5 Documents			
BM25	0.0431	0.1045	–
DigiLab	0.0433	0.1109	–
CiTIUS	0.0455	0.1119	–
WISE	0.0611	0.1198	BioBERT
WISE_{NLI}	0.0883	0.1823	BioBERT
GPT_{RAG}	0.1045	0.2098	BioBERT
$\text{Llama}_{\text{RAG}}$	**0.1079**	**0.2146**	BioBERT
$\text{Falcon}_{\text{RAG}}$	0.0994	0.2011	BioBERT
Top-10 Documents			
BM25	0.0784	0.1923	–
DigiLab	0.0823	0.1992	–
CiTIUS	0.0843	0.1999	–
WISE	0.1102	0.211	BioBERT
WISE_{NLI}	0.1302	0.2321	BioBERT
GPT_{RAG}	0.1502	0.2655	BioBERT
$\text{Llama}_{\text{RAG}}$	**0.1532**	**0.2702**	BioBERT
$\text{Falcon}_{\text{RAG}}$	0.1495	0.2568	BioBERT

TREC HM 2020 dataset

Model	CAM_{MAP}	CAM_{NDCG}	Embeddings
Top-5 Documents			
BM25	0.0631	0.1435	–
DigiLab	0.0712	0.1543	–
CiTIUS	0.0754	0.1554	–
WISE	0.0844	0.1608	BioBERT
WISE_{NLI}	0.0923	0.1922	BioBERT
GPT_{RAG}	0.1178	0.2234	BioBERT
$\text{Llama}_{\text{RAG}}$	**0.1222**	**0.2298**	BioBERT
$\text{Falcon}_{\text{RAG}}$	0.1123	0.2165	BioBERT
Top-10 Documents			
BM25	0.1047	0.2052	–
DigiLab	0.1186	0.2011	–
CiTIUS	0.1194	0.2095	–
WISE	0.1233	0.22	BioBERT
WISE_{NLI}	0.1341	0.2455	BioBERT
GPT_{RAG}	0.1547	0.2712	BioBERT
$\text{Llama}_{\text{RAG}}$	**0.1602**	**0.2723**	BioBERT
$\text{Falcon}_{\text{RAG}}$	0.1501	0.2665	BioBERT

A Tool for Explainability?



Can 5G antennas cause COVID-19?

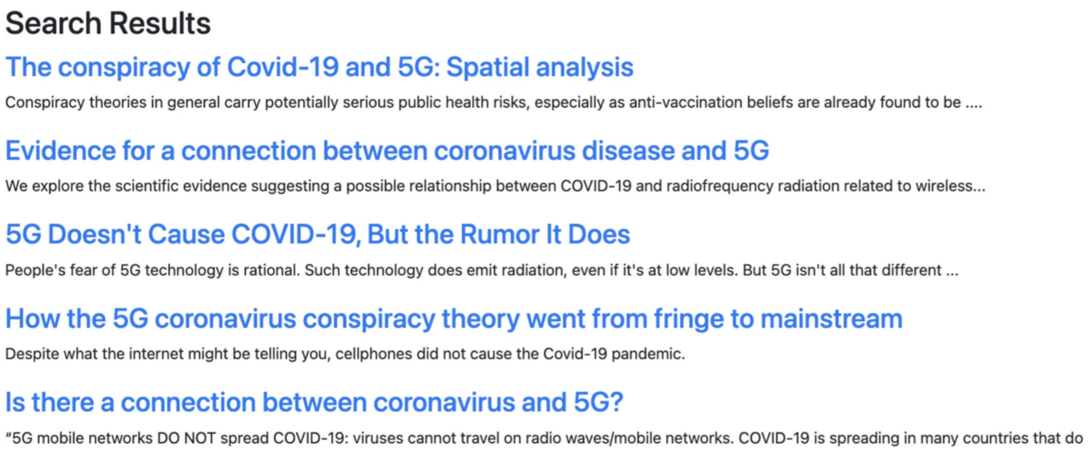
Search Results for "Can 5G antennas cause COVID-19"

LLM Generated Text

Based on the context provided, there is a misconception linking 5G antennas to the COVID-19 pandemic (Reference: 10316077). However, this connection has no statistically significant evidence to support it (Reference: 10316077). Instead, it's important to note that 5G networks play a crucial role in ensuring secure data handling and enhancing user privacy (Reference: 10255561). Moreover, SARS-CoV-2 variants remain the main cause of COVID-19 outbreaks (Reference: 10288941).

Reference List

- 10316077 - Softić, Adaleta, Elma Omeragić, Martin Kondža, Nahida Srabović, Aida Smajlović, Esmeralda Dautović, Nataša Bubić Pajić et al. "Knowledge and Attitudes regarding Covid-19 Vaccination among Medical and Non-medical Students in Bosnia and Herzegovina." Acta Medica Academica 52, no. 1 (2023): 1.
- 10255561 - Ullah, Yasir, Mardeni Bin Roslee, Sufian Mousa Mitani, Sajjad Ahmad Khan, and Mohamad Huzaimy Jusoh. "A survey on handover and mobility management in 5G HetNets: current state, challenges, and future directions." Sensors 23, no. 11 (2023): 5081.
- 10288941 - Soto, Ismael, Raul Zamorano-Illanes, Raimundo Becerra, Pablo Palacios Játiva, Cesar A. Azurdia-Meza, Wilson Alavia, Verónica García, Muhammad Ijaz, and David Zabala-Blanco. "A new COVID-19 detection method based on CSK/QAM visible light communication and machine learning." Sensors 23, no. 3 (2023): 1533.



Search Results

The conspiracy of Covid-19 and 5G: Spatial analysis

Conspiracy theories in general carry potentially serious public health risks, especially as anti-vaccination beliefs are already found to be ....

Evidence for a connection between coronavirus disease and 5G

We explore the scientific evidence suggesting a possible relationship between COVID-19 and radiofrequency radiation related to wireless...

5G Doesn't Cause COVID-19, But the Rumor It Does

People's fear of 5G technology is rational. Such technology does emit radiation, even if it's at low levels. But 5G isn't all that different ...

How the 5G coronavirus conspiracy theory went from fringe to mainstream

Despite what the internet might be telling you, cellphones did not cause the Covid-19 pandemic.

Is there a connection between coronavirus and 5G?

"5G mobile networks DO NOT spread COVID-19: viruses cannot travel on radio waves/mobile networks. COVID-19 is spreading in many countries that do

Outcomes and Discussion

Thank you for your attention | Grazie per l’attenzione

Some Bibliography

- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in neural information processing systems, 33, 9459-9474.
- Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., ... & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997, 2.
- Gao, Y., Xiong, Y., Wang, M., & Wang, H. (2024). Modular RAG: Transforming rag systems into lego-like reconfigurable frameworks. arXiv preprint arXiv:2407.21059.
- Huang, Y., & Huang, J. (2024). A survey on retrieval-augmented text generation for large language models. arXiv preprint arXiv:2404.10981.
- Fan, W., Ding, Y., Ning, L., Wang, S., Li, H., Yin, D., ... & Li, Q. (2024, August). A survey on RAG meeting LLMs: Towards retrieval-augmented large language models. In Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (pp. 6491-6501).
- Upadhyay, R., & Viviani, M. (2025). Enhancing Health Information Retrieval with RAG by prioritizing topical relevance and factual accuracy. Discover Computing, 28(1), 27.
- Locci, S., Audrito, D., Livraga, G., Viviani, M., & Di Caro, L. (2025). Leveraging RAG for Privacy Violation Detection and Explainability. Proceedings of the International Joint Conference on Neural Networks (IJCNN 2025), June 30-July 05, Rome, Italy