Generative AI for for Trend and Risk Analysis

Luca Cagliero, Giuseppe Gallipoli, Simone Papicchio, Lorenzo Vaiani

Politecnico di Torino, Turin, Italy

{name.surname}@polito.it

Politecnico di Torino

INTESA SANPAOLO INNOVATION CENTER

November 29th 2024, Turin, Italy

Input documents

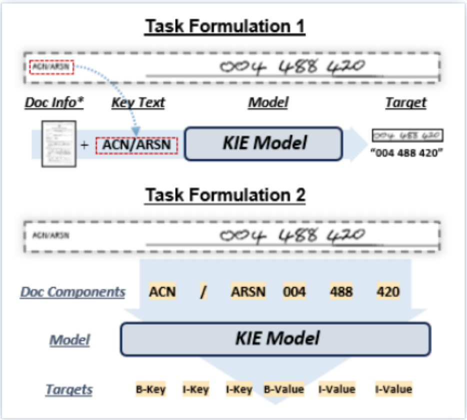
- Trend and Risk Analysis documents mainly consist of **Visually Rich Documents (VRDs)**
  - Long documents integrating various types of visually rich elements such as paragraphs, tables, charts, diagrams, and photos
  - Common format: PDF files (natively digital or scanned)
  - Elements are essential for illustrating, explaining, and summarizing information

Common VRD analysis

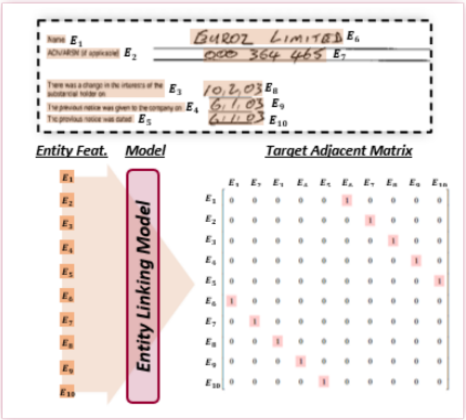
- **Key Information Extraction**
  - identify and extract values based on predefined keys

Sample Document

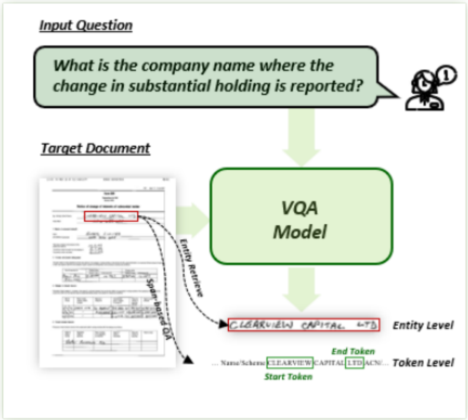
Key Information Extraction



Entity Linking

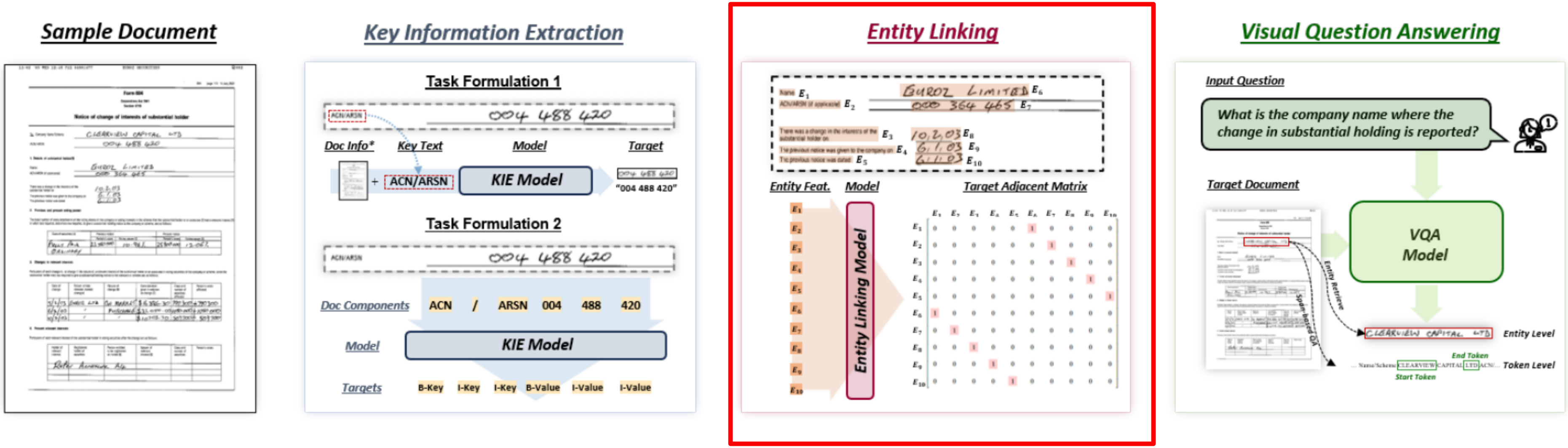


Visual Question Answering



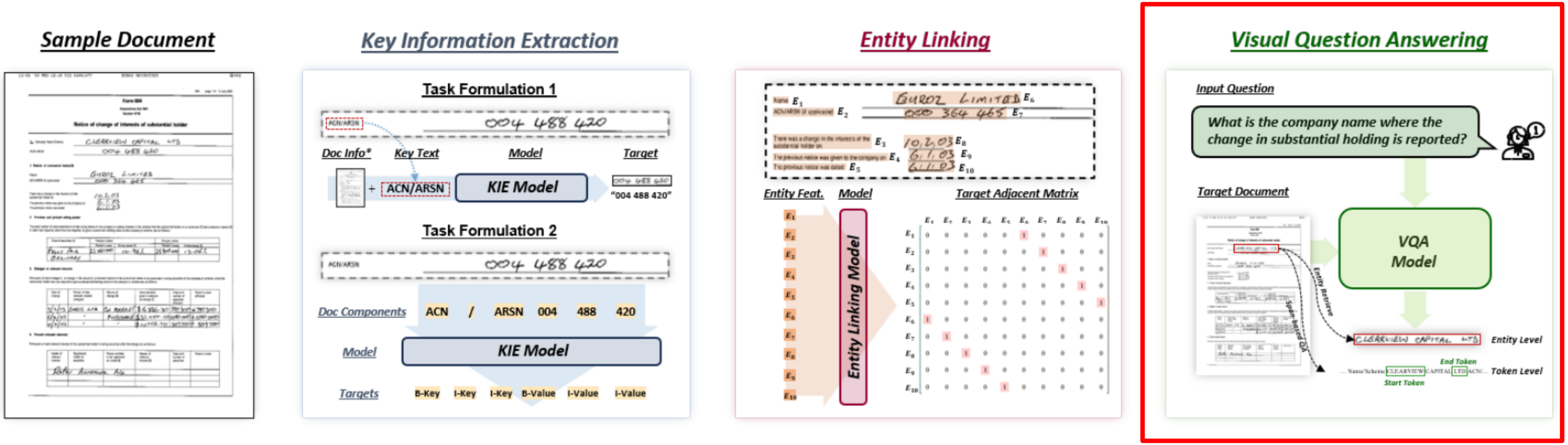
Common VRD analysis

- **Entity Linking**
    - identify the semantic correlations between document entities (parent-child relations)



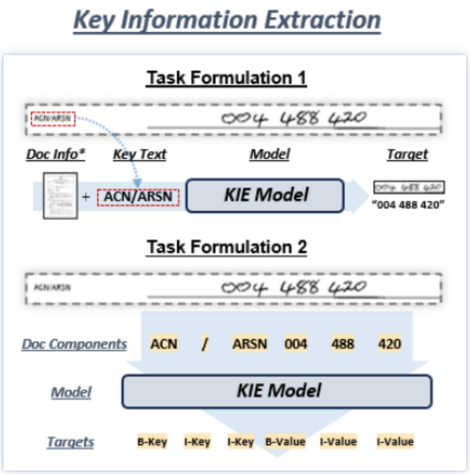
Common VRD analysis

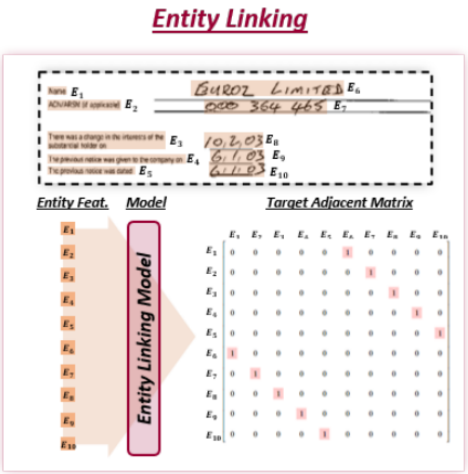
- **Visual Question Answering**
  - answering questions posed in natural language based on the contextual information in VRDs

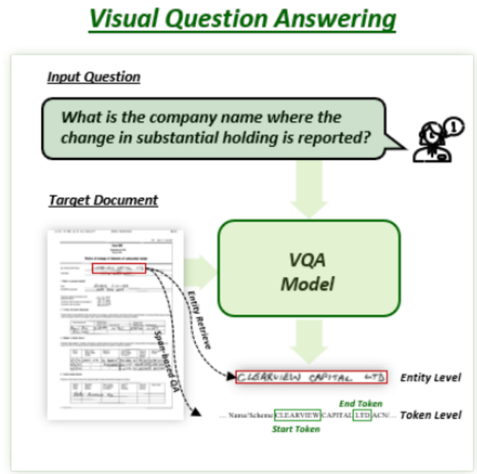


Common VRD analysis

- **Visual Question Answering + Summarization**
  - answering questions in natural language based on contextual information in VRDs
  - Condense the returned information into a query-driven summary

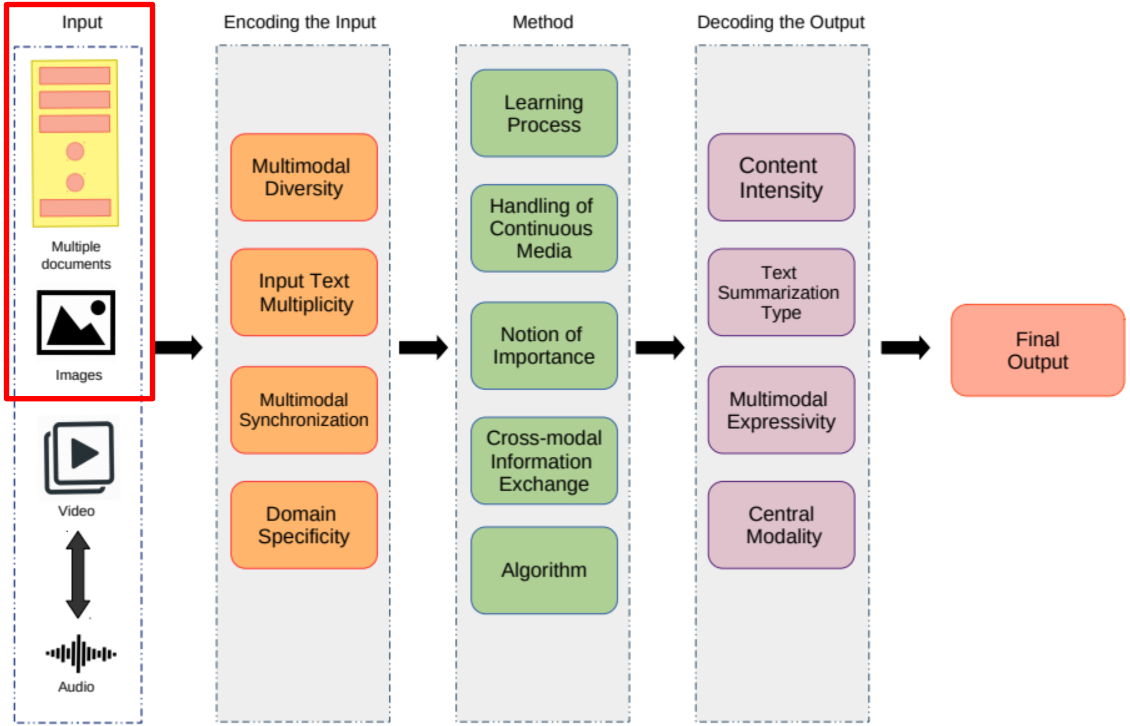






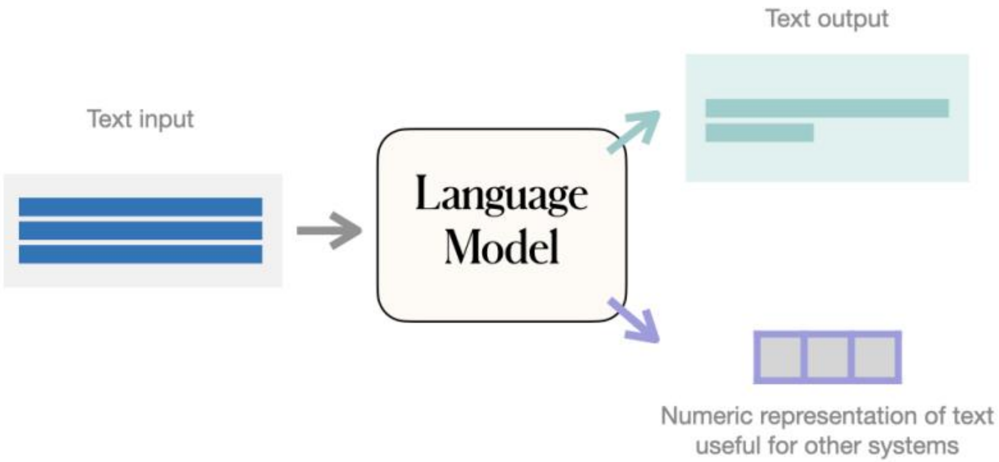
Multimodal Summarization

VRD sources



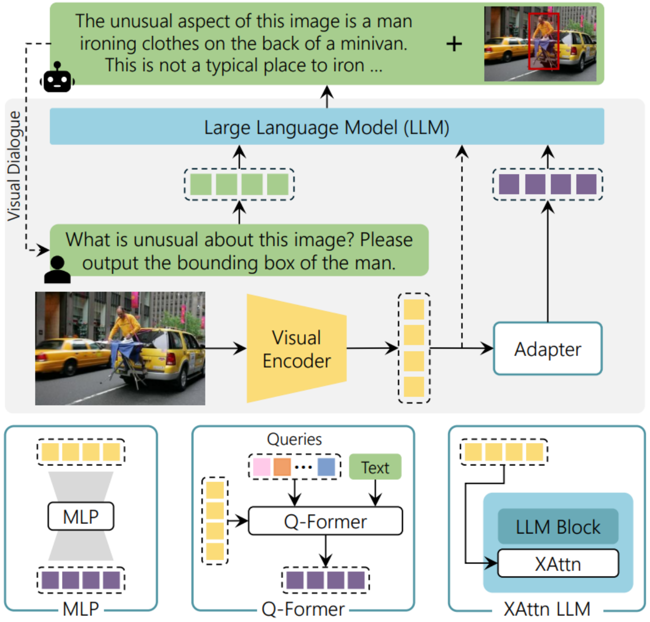
Classical LLMs

- Model the generative likelihood of word sequences to predict the probability of future (or missing) tokens
    - **Unsuited to Multimodal Content**



Multimodal LLMs

- Able to jointly reason obout textual and visual elements
  - **Suited to VRDs**



Multimodal reasoning using the Opensource LLM LLAvA

Multimodal LLMs

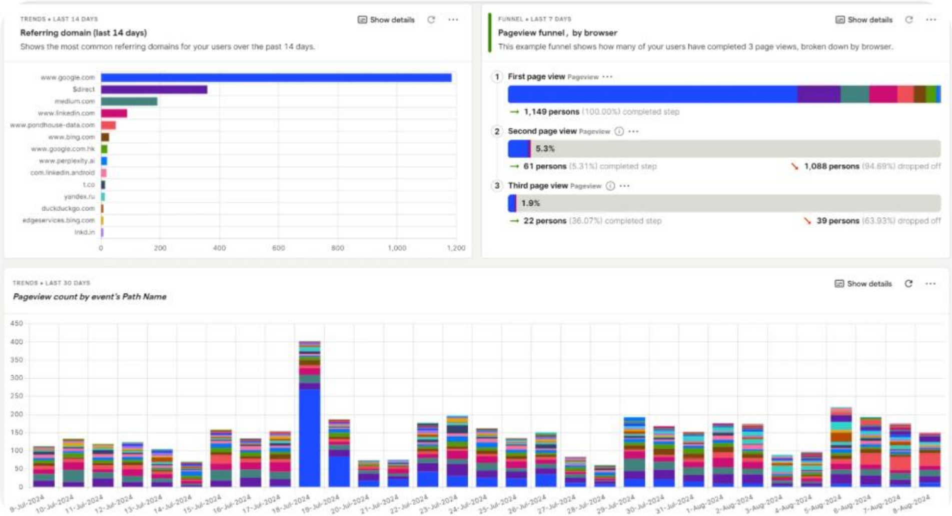
- Able to jointly reason on textual and visual elements
  - **Suited to VRDs**


```
with open("./public/blog/gpt4ovision/dashboard.png", "rb") as image_file:
  base64_image = base64.b64encode(image_file.read()).decode('utf-8')

messages=[
  {
    "role": "system",
    "content": """You are a data analyst. You get an image of a dashboard.
    Create a written report based on this dashboard.
    The report should be short and concise and should contain the most important metrics.
    The report is aimed at high-level management, so make sure it only covers the most prevalent facts.
    Focus on actionable insights.
    """
  },
  {
    "role": "user",
    "content": [
      {"type": "text", "text": "Create a written summary of this dashboard."},
      {
        "type": "image_url",
        "image_url": {
          "url": f"data:image/jpeg;base64,{base64_image}",
          "detail": "high"
        },
      },
    ],
  }
]
```

Image reasoning using the proprietary GPT4o

Multimodal LLMs

- Able to jointly reason on textual and visual elements
    - **Suited to VRDs**



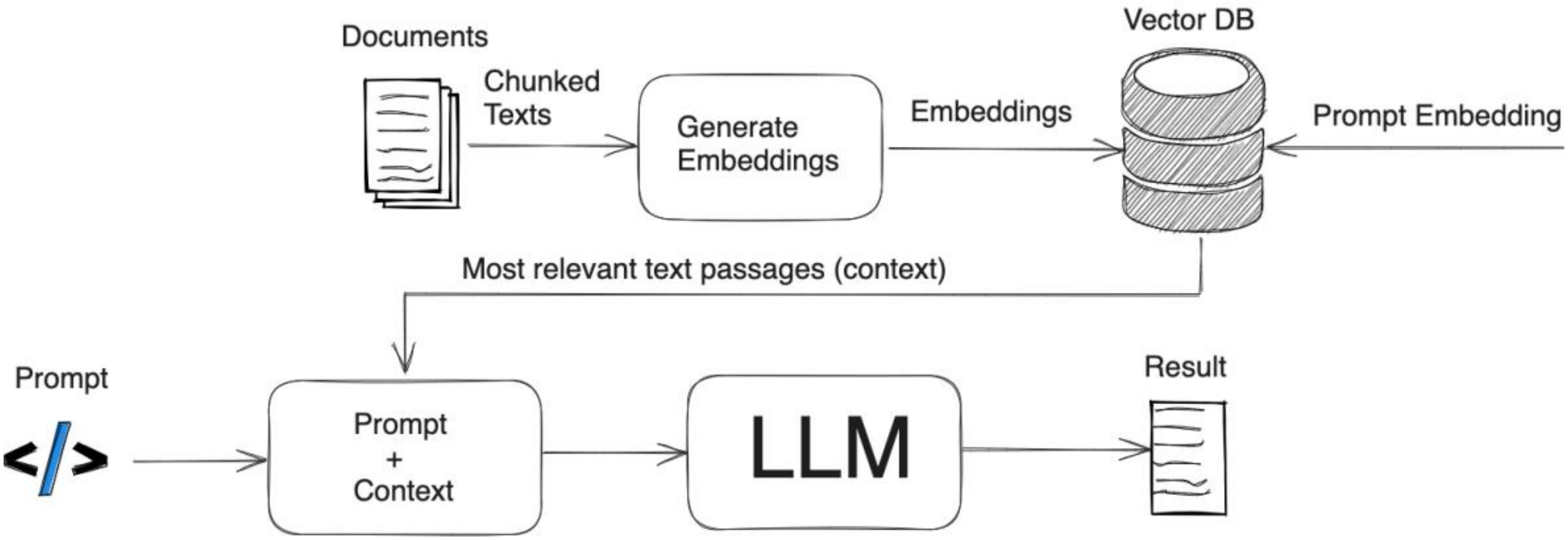

```
### Dashboard Summary Report

#### Referring Domains (Last 14 Days)

- **Top Referring Domain:** www.google.com is the leading referrer with approximately 1,100+ referrals.
- **Direct Visits:** The second highest source, with around 400 direct visits.
- **Other Key Referrers:** Medium.com, linkedin.com, and pondhouse-data.com also contribute notably to the referral

#### Pageview Funnel (Last 7 Days)

- **First Page View:** 1,149 users (100%) completed the first page view.
- **Second Page View Drop-off:** 5.3% or 61 users continued to the second page view, indicating a 94.69% drop-off (
- **Third Page View Drop-off:** From the second page, only 1.9% or 22 users progressed to the third page view, with

#### Pageview Count by Event's Path Name (Last 30 Days)

- **Daily Variation:** Pageview counts exhibit significant daily fluctuations, generally ranging between 50 to 300
- **Peak Traffic Day:** Notably, July 18, 2024, saw a peak in pageviews, surpassing 400 views, indicating a potenti

### Actionable Insights

1. **Enhance Engagement Post First Page:** The significant drop-offs after the first page view indicate potential i
   Further analysis and improvement of the user experience after the first page could help retain users.
2. **Focus on Key Referrers:** Investment in strengthening relationships, collaborations or advertising with top re

3. **Investigate High-traffic Days:** The spike in traffic on July 18, 2024, warrants closer inspection to understa

4. **Continual Monitoring and Testing:** Regular monitoring is essential to understand trends better and identify s
```

Image reasoning using the proprietary GPT4o

Known LLM issues

- Bias
- Hallucination
- Temporal updates
- Math word problems
- Privacy of data exchanged with LLM providers
- …

Solution: Retrieved Augmented Generation



Research Question #1

- Given a topic of interest, analysts are particularly interested in retrieving the top trending keywords related to it and then use them to annotate the most relevant document elements (e.g., text paragraphs, images or tables).

Problems' statement

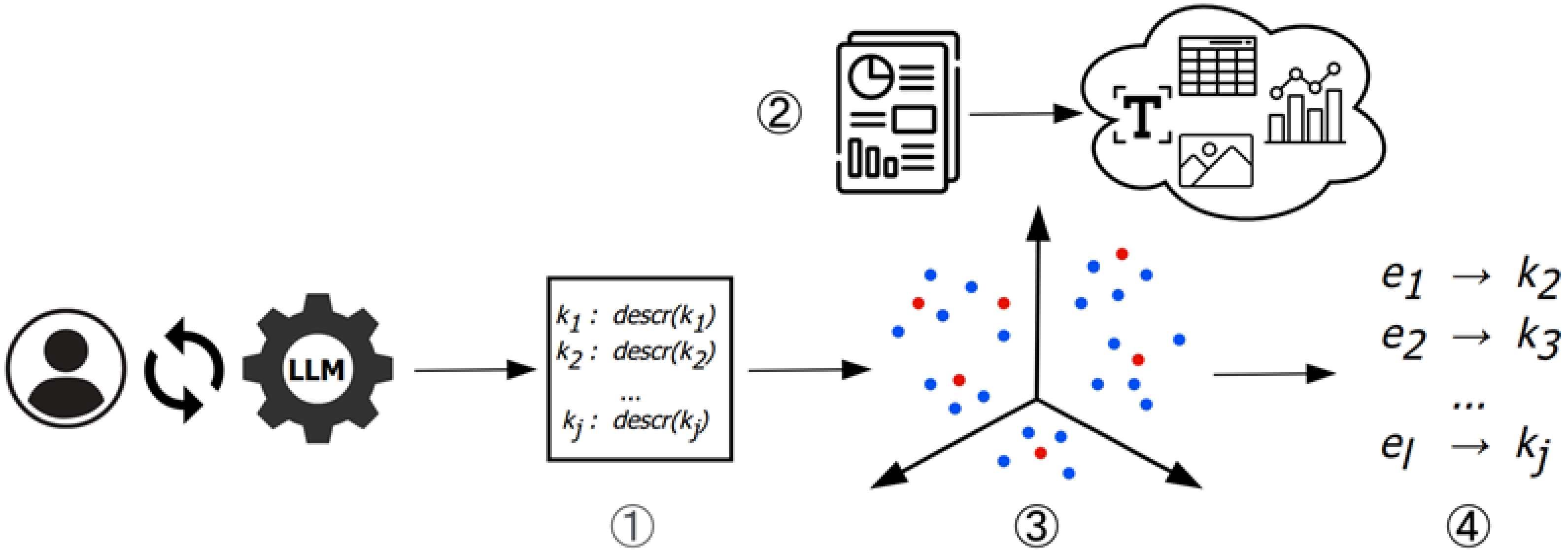
Given a set of multi-page financial documents $\mathcal{D}$ and a set of facets $\mathcal{F}$ describing the topics of interest, our purpose is threefold:

1. **Keyword generation and description**: Generate for each facet $f_i \in \mathcal{F}$ a set of keywords $k_j \in \mathcal{K}^i$ related to f_i. Next, annotate each keyword k_j with a free-text description $descr(k_j)$ summarizing its meaning.

1. **Captioning of non-textual document elements**: Produce textual descriptions of multimedia document elements $e_l \in \mathcal{E}^m$, where an arbitrary element e_l in a document $d_m \in \mathcal{D}$ can be either an image, a table, or a textual paragraph.

1. **Keyword-based content annotation**: For each element e_l, retrieve the keywords k_j that are most relevant to e_l.

Keyword-based document annotation pipeline



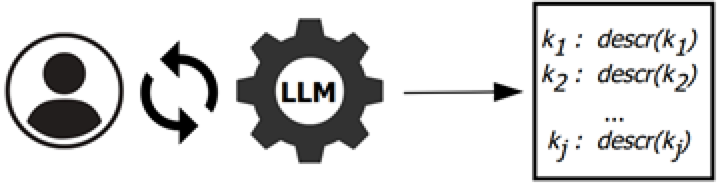
① Keyword and description generation

② Document pre-processing

③ Document element and keyword description encoding

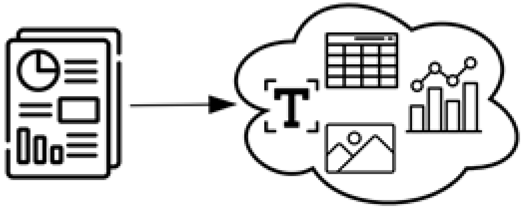
④ Keyword-based content annotation

Step 1 – Keyword and description generation



- Given a facet name provided by the expert, we use the LLM to automatically generate a set of related keywords as well as the corresponding free-text descriptions.
- We explore different settings:
  - **Zero-Shot learning** – *Cold Start*: Prompt the LLM with the facet name only.
  - **Few-Shot learning** – *Cold Start*: Prompt the LLM with the facet name and examples chosen randomly of keywords and their corresponding descriptions.
  - **Few-Shot learning** – *Additional Keyword Recommendation*: Prompt the LLM with the facet name and examples of facet-related keywords and their corresponding descriptions.

Step 2 – Document pre-processing



- We extract the following three main elements:
  (i) **Textual paragraphs**, (ii) **Visual items**, and (iii) **Tables**.
- Paragraphs and tables are extracted using Azure Document Intelligence service [1].
- For visual and textual content extraction, we face the following challenges:
  - *Slide extraction*: Slides are unsuitable for text and image extraction using standard tools. To address this issue, we train a CNN classifier to detect if the current page is a slide. In this case, we generate a textual explanation of the slide content using GPT-4 Vision [2].
  - *Paragraph length*: We filter out extracted textual elements consisting of less than 4 words.
  - *Redundant table content*: We filter out duplicated text between tables and paragraphs.
  - *Irrelevant images*: We filter out irrelevant visual items using some heuristics.

Example

- *Dataset*: ICT Risk Analysis
- *Facet*: cyber risk
- *Keyword*: third-party risk
- *Reference description*: “It refers to the potential risks or threats to an organization arising from relationships with third parties, such as suppliers, business partners or external [...]”
- *Generated description*: It is the risk that arises from the use of third-party vendors, suppliers, or partners that provide goods or services to an organization. Third-party [...]”
- *Document element*: “The image presents [...] in the context of retail banking leaders. [...] security providers aim to protect company, payment, card, and consumer data [...] the importance of various data privacy and security [...]”
- *Target keywords* : third-party risk, regulation
- *Assigned keywords*: third-party risk, regulation, compliance

Dataset & Evaluation Metrics

- Business Units provided the following three datasets:
  - **ICT Risk Analysis**: 11 documents, annotated with 2 facets and 25 keywords. It contains 991 textual elements, 13 images, and 15 tables.
  - **Trend Analysis**: 4 documents, annotated with 1 facet and 12 keywords. It contains 69 images (mostly presentation slides).
  - **Innovation Analysis**: 3 documents, annotated with 3 facets 12 and 19 keywords. It contains 82 textual elements, 32 images, and 18 tables.

To evaluate the efficacy of our pipeline, we employ the following metrics:

- **Keyword and description generation**:
  - ROUGE-1/2/L F1-score (R1/2/L)
  - BERTScore F1-score (BS)
- **Document element annotation**:
  - Precision at K (P@K)
  - Recall at K (R@K)
  - Mean Reciprocal Rank (MRR)

Research Question #1 - Keyword and description generation results

- Dataset: ICT Risk Analysis

	K = unspecified		$K=3$		$K=5$		$K=10$		$K=20$	
	GPT-4	Llama2	GPT-4	Llama2	GPT-4	Llama2	GPT-4	Llama2	GPT-4	Llama2
RL	0.051	**0.066**	**0.070**	0.062	0.058	**0.062**	0.057	**0.065**	0.058	**0.060**
BS	**0.860**	0.859	0.862	**0.864**	0.861	**0.863**	**0.865**	0.860	**0.861**	0.857
P@K	0.771	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	0.867	**0.944**	0.833	**0.894**
R@K	**0.447**	0.296	**0.133**	**0.133**	**0.221**	**0.221**	0.375	**0.420**	0.721	**0.783**

Keyword generation for varying K. English language.

	Italian		English	
	Camoscio	GPT-4	Llama2	GPT-4
R1	0.310	**0.413**	0.394	**0.437**
R2	0.082	**0.169**	0.131	**0.150**
RL	0.208	**0.279**	0.254	**0.284**
BS	0.719	**0.773**	0.760	**0.902**

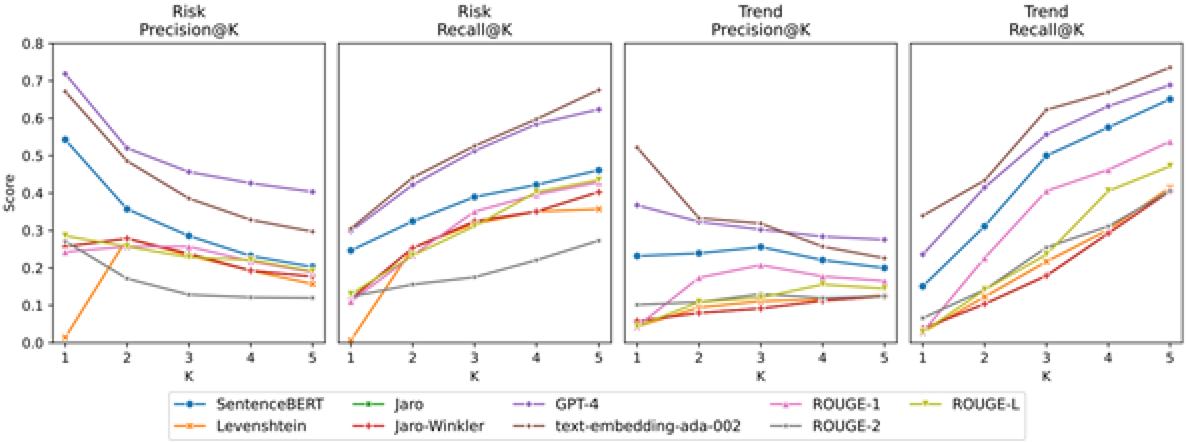
Keyword description generation.

	Italian	English
Usefulness	**4.38**±1.60	4.33±1.78
Coherence	4.52±0.93	**4.62**±1.15
Non-Redundancy	4.38±1.12	**4.52**±1.11
Grammaticality	4.60±0.73	**4.81**±1.20
Overall Quality	4.33±1.37	**4.34**±1.66

Human evaluation of keyword descriptions.

- **Keyword generation**: open-source LLMs are highly competitive.
- **Keyword description generation**: GPT-4 performance is superior to that of open-source models (Llama2, Camoscio) for both languages.
- Better performance on English than Italian texts for both LLMs.
- Prompting LLMs with few examples is beneficial for both subtasks.
- **Human evaluation** of keyword descriptions involving domain experts. Results are satisfactory and coherent with quantitative metrics.

Research Question #1 – Document Annotation Results



Similarity measure	ICT Risk Analysis	Trend Analysis
R1	0.458	0.300
R2	0.367	0.279
RL	0.472	0.258
Levenshtein	0.347	0.247
Jaro	0.483	0.249
Jaro-Winkler	0.483	0.249
SentenceBERT	0.658	0.430
embedding-ada-002	**0.779**	**0.610**
GPT-4	0.729	0.500

Mean Reciprocal Ranks.

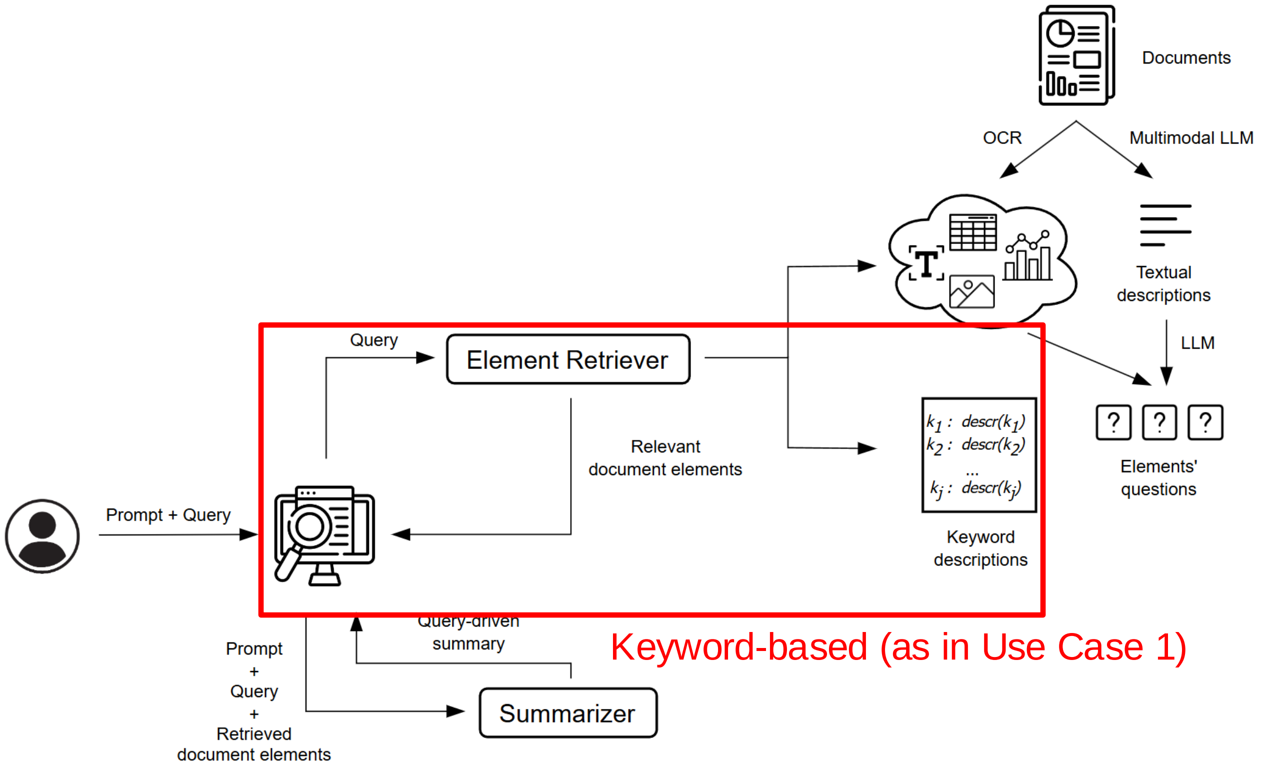
- Best performance achieved by textual **semantic similarity** based on contextual embeddings and LLM prompting.
- Similarity based on OpenAI embeddings performs best.
- While increasing the number K of retrieved keywords, precision decreases whereas recall increases.

Research Question #2

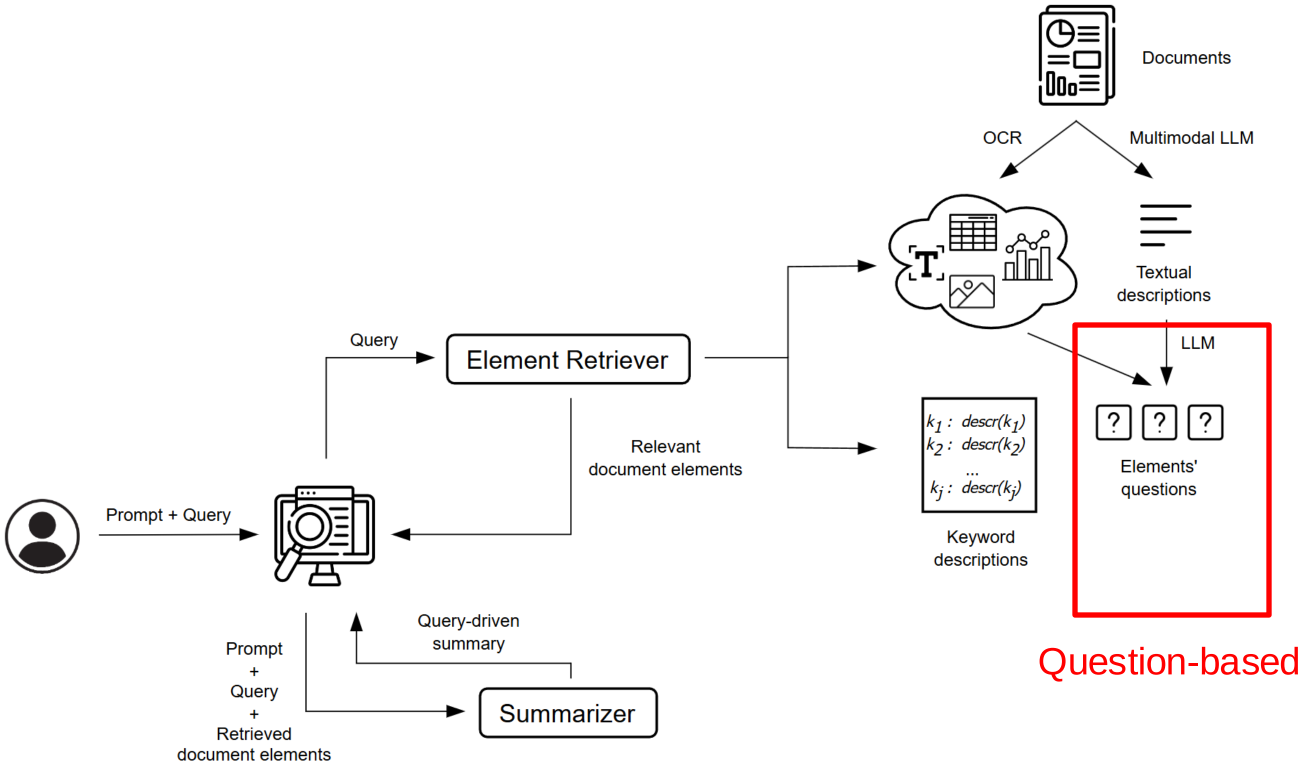
- To gain insights into rapidly evolving, complex topics, TRA analysts are interested in posing **natural language questions** on VRD content and generate **query-driven summarized answers**.

Query-driven Summarization of Visually Rich Documentsto support Trend and Risk Analysis in Banking.
Giuseppe Gallipoli, Simone Papicchio, Lorenzo Vaiani, Luca Cagliero, Arianna Miola and Daniele Borghi.
Currently under review at Elsevier Computer In Industry Journal ISSN: 0166-3615

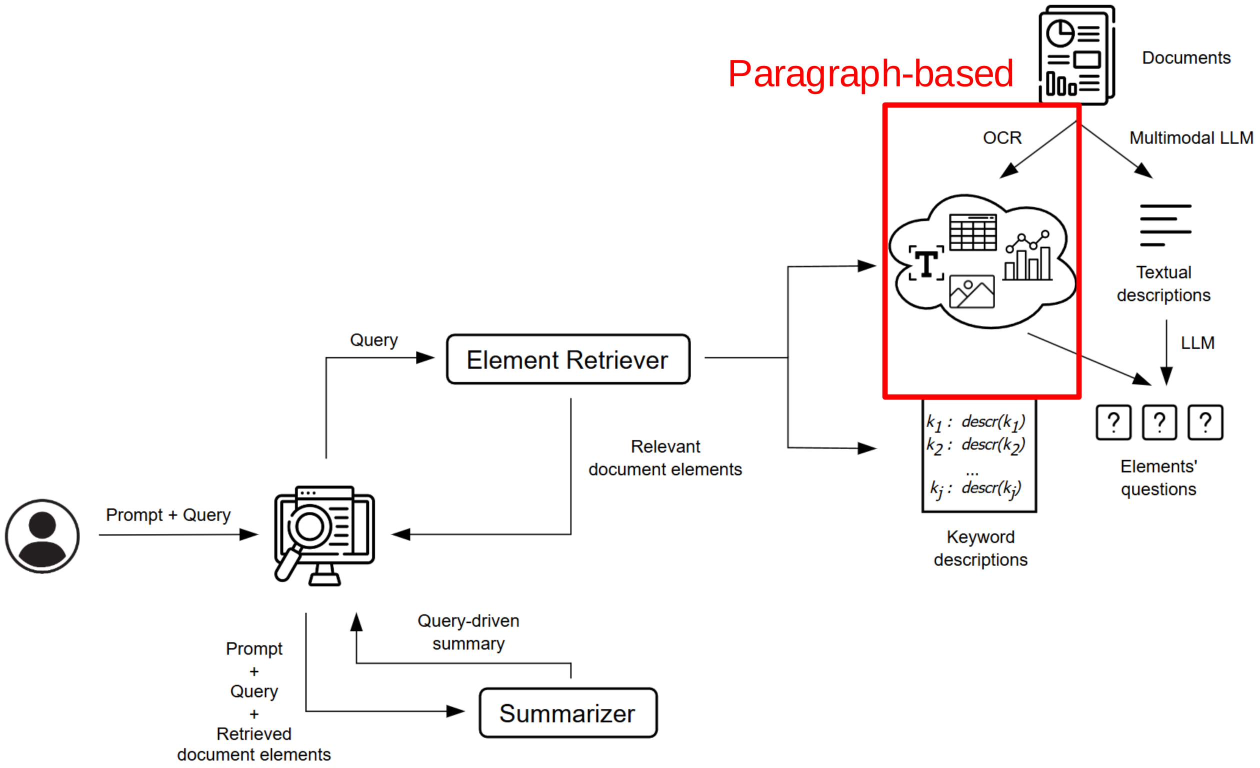
Integrated RAG systems



Integrated RAG systems

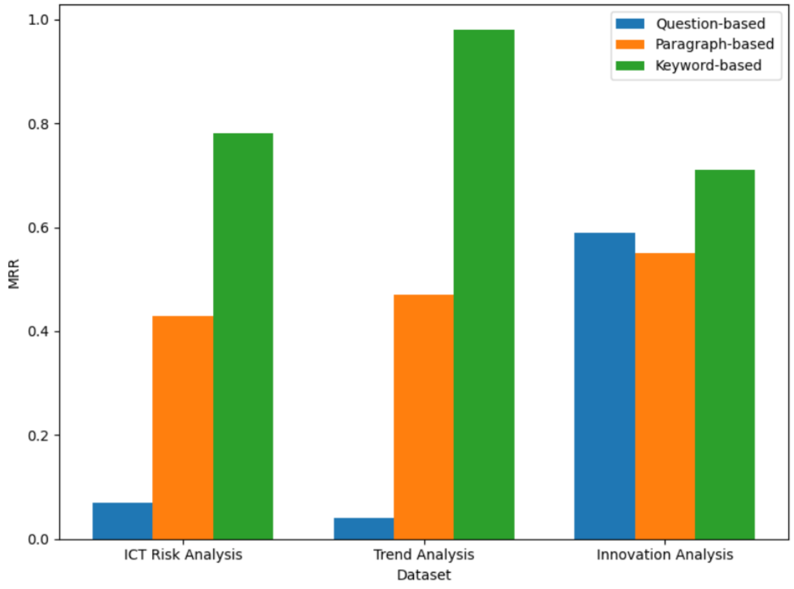


Integrated RAG systems



Research Question #2 – Main results

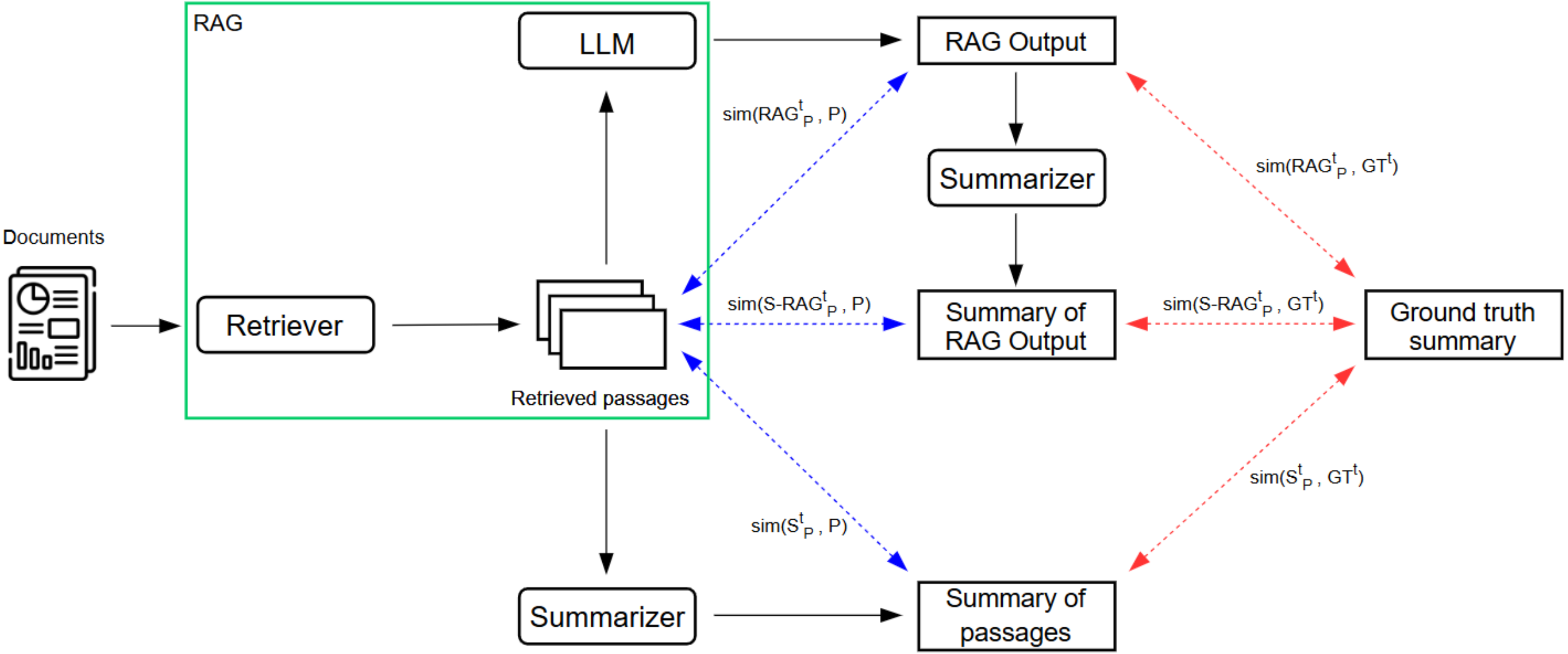
- The summarization performance on VRDs such as Trend documents are comparable to those obtained on more traditional document types such as Innovation and Risk.
- The open-source LLM Llama3-Instruct achieves performance comparable to those of the proprietary GPT-4o model on English-written Innovation and Risk documents whereas the gap in performance between commercial tools and open-source solutions is still significant on Italian documents.
- The Keyword-driven element retrieval strategy achieves retrieval performance superior to paragraph- and question-based ones.



Research Question #3

- Given a question, the RAG system returns (1) The retrieved elements (i.e., the passages), (2) The answer (in plain text), and (3) A separate query-driven summary.
- Which outputs should the TRA analyst inspect (first)? To what extent are the outputs different?

Experimental design



Research Question #3 - Main results

- The **answers** provided by **Classical RAG** to TRA-related questions are inherently **redundant**, calling for ad hoc summarization strategies.
- While proprietary LLMs (GPT4o) excel at generating concise summaries of the retrieved passages, **the level of synthesis of open-source models** (including LLMs) is, in general, **not satisfactory**.
- **Cascading RAGs** with summarization modules (regardless of the approach used) is **not beneficial**.

[9]

Thank you for your attention!

Politecnico di Torino

INTESA SANPAOLO INNOVATION CENTER