



MedPix 2.0: A Comprehensive Multimodal Biomedical Data Set for Advanced AI Applications with Retrieval Augmented Generation and Knowledge Graphs

Irene Siragusa¹ · Salvatore Contino¹ · Massimo La Ciura¹ · Rosario Alicata¹ · Roberto Pirrone¹

Received: 12 December 2024 / Revised: 30 April 2025 / Accepted: 3 May 2025
© The Author(s) 2025

Abstract

The increasing interest in developing Artificial Intelligence applications in the medical domain, suffers from the lack of high-quality data set, mainly due to privacy-related issues. In addition, the recent increase in Vision Language Models (VLM) leads to the need for multimodal medical data sets, where clinical reports and findings are attached to the corresponding medical scans. This paper illustrates the entire workflow for building the MedPix 2.0 data set. Starting with the well-known multimodal data set MedPix[®], mainly used by physicians, nurses, and healthcare students for Continuing Medical Education purposes, a semi-automatic pipeline was developed to extract visual and textual data followed by a manual curing procedure in which noisy samples were removed, thus creating a MongoDB database. Along with the data set, we developed a Graphical User Interface aimed at navigating efficiently the MongoDB instance and obtaining the raw data that can be easily used for training and/or fine-tuning VLMs. To enforce this point, in this work, we first recall DR-Minerva, a Retrieve Augmented Generation-based VLM model trained upon MedPix 2.0. DR-Minerva predicts the body part and the modality used to scan its input image. We also propose the extension of DR-Minerva with a Knowledge Graph that uses Llama 3.1 Instruct 8B, and leverages MedPix 2.0. The resulting architecture can be queried in an end-to-end manner, as a medical decision support system. MedPix 2.0 is available on GitHub.

Keywords MedPix · Biomedical data set · VLM · Decision Support System

1 Introduction

The rise of computer-based applications in recent years has strongly favored the digitization of historically analogous processes in the management and analysis of biomedical data. In turn, the emergence of technologies based on Artificial Intelligence (AI) prompted the development of increasingly precise models to support diagnosis for the construction of personalized treatments. In the biomedical domain, in fact, different subareas can benefit from AI-based systems, ranging from the administrative field (queue management in emergency areas, centralized integration of EHRs, and so on) to the clinical one, thanks to the use of AI that can efficiently extract useful features to achieve diagnosis reliably.

One of the fundamental requirements of these applications is their trustworthiness, as they must help physicians with a high degree of confidence, being able to provide reliable predictions and classifications. However, AI models require a considerable amount of data to achieve good performance. Data availability makes expansion into

Irene Siragusa and Salvatore Contino have contributed equally to this work.

✉ Salvatore Contino
salvatore.contino01@unipa.it

Irene Siragusa
irene.siragusa02@unipa.it

Massimo La Ciura
massimo.laciura@community.unipa.it

Rosario Alicata
rosario.alicata@community.unipa.it

Roberto Pirrone
roberto.pirrone@unipa.it

¹ Department of Engineering, University of Palermo, Viale delle Scienze, Palermo 90128, Sicily, Italy

the biomedical domain much more complex than in other ones. One of the main problems lies in the availability of data sets to allow the scientific community to develop new AI approaches. This problem arises mainly from the sensitive nature of data that encompasses privacy issues, and this makes it difficult to build public data sets containing images and/or clinical reports to be available to the scientific community.

To overcome these obstacles, the European Community created the European Health Data Space (EHDS).¹ This is a health-specific ecosystem composed of common rules, standards and practices, infrastructure, and a governance framework that aims to empower people through increased digital access and control of their personal electronic health data. The EHDS promotes a single market for electronic health record systems, relevant medical devices, and high-risk artificial intelligence systems. Finally, the EHDS aims to provide the researchers who make use of health data with a trusted framework to control the entire analytical process [1, 2]. The systemic approach proposed by the EHDS will provide a controlled pool of both shared data and applications. It will allow AI in the medical field to access certified and controlled data by overcoming both medical and engineering obstacles through new solid foundations on which a new generation of AI-based health applications can rely.

In view of the implications provided by the EHDS on the application side, researchers wishing to pursue their activity in the biomedical domain by developing Vision Language Models (VLMs) should be able to find and optimize the data sets that are available in a way that maximizes the available public resources. To date, there are not so many data sets that contain images, such as Computed Tomography (CT) and Magnetic Resonance (MR)² scan, and medical reports. One of the most well-known data sets is MedPix[®],³ a free open-access online database of medical images, teaching cases, and clinical topics, integrating images and textual metadata provided by the National Library of Medicine (NLM), the largest collection of medical documentation in the world. This collection is part of the much more famous National Institutes of Health, which is responsible for managing health and biomedical research in the United States, actively participating in the identification of new drugs from 2010 to 2016 in collaboration with the Federal Drug Administration. Given its important contribution to science, this institution is very careful about managing the privacy issue and applies strict rules for anonymising samples stored in their repositories [3, 4], which have also been applied to

the MedPix data set used as input in the proposed work. MedPix[®] includes more than 12,000 patient case scenarios, 9,000 topics, and nearly 59,000 images. The contribution proposed in this paper lies in both curation and a new arrangement of MedPix[®] data in a non-relational database based on MongoDB. This new arrangement of the MedPix[®] structure can be regarded as a brand new data set that we call *MedPix 2.0*. MedPix 2.0 makes original data accessible and well structured, as it is very easy to build views capable of creating ready-made subsets for training VLMs. Database queries have been further simplified by developing a user-friendly Graphical User Interface (GUI). To enforce the usability of MedPix 2.0, the user is made able to pose her/his query, and browse the results in a very close way as in the original website. For improved usability of MedPix 2.0, the user can formulate their own question and browse the results in a way that is very similar to the original website. In fact, unlike the original MedPix, which was created as an educational platform, our data set allows you to dynamically build training and/or test data sets of various types (image-only, text-only or multimodal) for AI models. The use of a NoSQL database that offers a schema-on-read approach to data allows diversifying access to our data set by any downstream application, in order to simplify the querying of the data set. In particular, it is extremely simple to construct documentary data sets and prompt instructions or few shots for a downstream LLM. Lastly, since we started from a data set containing documents that were not structured but simply organised into sections, it seemed natural to use a loose structuring that reflected the organisation of the original data.

Another contribution of this work is a direct demonstration of how MedPix 2.0 makes it easy to train and test VLMs. At first, we recall the architecture and the training phase of DR-Minerva [5], a VLM developed by the authors that leverages the Minerva LLM [6] and uses Flamingo [7] for multi-modality and that exploits the Retrieval Augmented Generation (RAG) approach [8]. RAG-based models are spreading in diverse domains, including the bio-medical one, reaching promising results with a retrieval module composed of both vector stores [9, 10] and graph structures [11]. We used our MongoDB data source to generate the train/validation/test split used for DR-Minerva.

Then, we propose a Knowledge Graph (KG) built from MedPix 2.0, using Llama 3.1 Instruct 8B [12]. The KG can be queried to obtain information about medical diseases and their diagnoses. The entire system, that is, the KG coupled with DR-Minerva, can be queried end-to-end, starting from a medical image to obtain suggestions about the most probable diagnosis with a free-text answer, thus working as a medical decision support system.

The paper is arranged as follows: Sect. 2 illustrates the State Of The Art (SOTA) for medical multimodal data sets.

¹ <https://www.european-health-data-space.com/>.

² We will refer also to MR scans as MRI, Magnetic Resonance Imaging.

³ <https://medpix.nlm.nih.gov/home>.

The details on the building process of the curated data set are reported in Sect. 3 along with the implementation of the MongoDB database and the GUI. The experimental results, the architecture of DR-Minerva, and the creation of the KG are reported and discussed in Sect. 4, while Sect. 5 illustrates the experimental results. Concluding remarks and future works are drawn in Sect. 7.

2 Related Works

Medical data sets for AI applications suffer from diverse problem, related both to the data and the peculiarity of the domain. First, there is a privacy issue, since clinical data contain private information about the patient, thus the process of creation of such type of data set has to start with an anonymization phase. To overcome this problem, researchers rely on either open-access and textbook data, or they collaborate with hospitals to create data sets. The first two methods allow for large scalability, relying on anonymous data. On the other hand, anonymization must be carried out from scratch when dealing with hospitals. Moreover, multimodal medical data suffer from scarcity when compared with other data sets related to different domains, such as MS-COCO [13].

One of the most used open-access multimodal databases for developing medical data sets is PubMed Central® (PMC).⁴ This is a widely used free archive of biomedical scientific literature, from which it is possible to build one's own data sets via semi-automatic procedures. In PMC, data are anonymized, and high-quality captions can be extracted from the medical research papers the images belong to. The following multimodal data sets were extracted from PMC: ROCO [14], MedICaT [15], PMC-OA [16]. ROCO contains pairs of radiology images and the corresponding captions, and it incorporates an out-of-class set to improve prediction and classification performances. MedICaT is a disjoint data set from ROCO that is mainly composed on radiology images and provides manually annotations for sub-figures. PMC-OA is the larger than the previous ones, and it keeps a variety of diagnostic procedures, diseases, and findings, while introducing sub-figure separation.

VQA-RAD [17] is a data set derived from MedPix®, and it collects a subset of radiological images, while providing Question-Answer (QA) pairs validated by domain experts.

Another source of available high-quality data is textbooks: PathVQA [18] is a Visual Question Answering (VQA) data set that collects both closed- and open-ended

QA pairs, which are extracted from both pathology textbooks and online digital libraries via a semi-automated pipeline.

On the other side, open access data sets like MIMIC-CXR [19], IU-Xray [20] and SLAKE [21] are manually annotated by domain experts. Both MINIC-CXR and IU-Xray are chest radiography data sets derived from hospital's clinical cases. Both if they contain a semi-structured radiology report, describing the radiological findings of the images it is related to. SLAKE, on the contrary, collects images from different radiology open-access data sets and provides manual annotations and QA pairs given by experienced doctors in English and Chinese.

INSPECT [22] is a worth mentioning multimodal data set that collects computed tomography pulmonary angiography (CTPA), radiology reports and structured data from EHRS, for pulmonary embolism diagnosis.

The aforementioned data sets are multimodal ones, and they are deeply focused on VQA tasks. Also some specialized data sets exist that provide only visual data, as UniToChest [23], ADNI [24], OMI-DB [25] or OASIS-3 [26] suitable for image segmentation and classification, or free-form textual data such as in Named Entities in Medical Case Reports (NE-MCR) in which the main focus is towards Named Entity Recognition (NER) [27]. Other works worth mentioning are UniToBrain [28] that integrates a technical report with the scanning modality, and the E3C corpus [29] that is an annotated multilingual data set of clinical reports suitable for NER and Relation Extraction (RE) tasks.

The proposed MedPix 2.0 data set may be close to VQA-RAD, they differ in both the sampling strategy and the samples themselves. In Med Pix 2.0, we do not integrate QA pairs, but different QA pairs can be created relying on the structured textual information provided in the data set itself, following the structure of the other one. Thus, the creation of more complex tasks is possible, such as document summarization or understanding.

With respect to the existing multimodal data sets, MedPix 2.0:

1. is derived from a freely open-access source, and it has no privacy-related issues;
2. offers a balanced variety of CT and MRI scans of different body parts;
3. for each image, a complete structured clinical case is provided.

Thanks to the annotation scheme we selected for the JSON documents in MedPix 2.0, the last point makes it suited for various tasks not only limited to document-level retrieval. Unfortunately, images are provided in PNG format, which limits visual processing with respect to raw images in

⁴ <https://www.ncbi.nlm.nih.gov/pmc/>.

DICOM format. A comprehensive overview of the reported data set, is reported in Table 1.

3 MedPix 2.0

MedPix® is a free open-access multimodal online database of medical images, teaching cases, and clinical topics, managed by the National Library of Medicine (NLM) of the National Institutes of Health (NIH). It mainly serves as a support system for Continuing Medical Education (CME) of physicians, nurses, and healthcare students. The database collects clinical cases related to more than 12,000 patients.

Each case contains at least one medical image, and the corresponding findings, discussion notes, diagnosis, differential diagnosis, treatment, and follow up. Textual information is reported in a semi-structured format. Attached to the clinical case, there is the topic section, where the disease under investigation is discussed in detail from an academic and general perspective.

In Fig. 1 an example from the MedPix® website is reported.

Despite the richness of the data set, its free availability, the possibility to add new cases, and the access to clinical cases of interest using wither keywords, the body part or the disease, it is not possible to access the raw data. This feature limits the usage of MedPix® for training multimodal

Table 1 Data set taxonomy

Data set	References	# samples	Data source	Task
<i>Textual-only</i>				
NE-MCR	[27]	53	PMC	NER
E3C	[29]	10.034	PMC	NER, RE
<i>Visual-only</i>				
UniToBrain	[28]	258	Hospital	Image segmentation
UniToChest	[23]	306.440	Hospital	Image segmentation
OASIS-3	[26]	6.471	Hospital	Image segmentation
OMI-DB	[25]	3.072.878	Hospital	Image classification
ADNI	[24]	1921	Hospital	Image classification
<i>Visual and textual</i>				
ROCO	[14]	87.952	PMC	Multimodal classification
MedICaT	[15]	217.060	PMC	Subfigure-subcaption alignment
PMC-OA	[16]	1.650.00	PMC	Multimodal classification
VQA-RAD	[17]	3.515	MedPix®	VQA
MINIC-CXR	[19]	377.110	Hospital	Classification, text generation
IU-Xray	[20]	8.121	Hospital	Multimodal retrieve
SLAKE	[21]	14.028	Open data	VQA
PathQA	[18]	32.799	Pathology books	VQA
INSPECT	[22]	23.248	Hospital	Multimodal classification
MedPix 2.0		2050	MedPix®	Multimodal classification, text generation

Fig. 1 An example from the MedPix® web site where the original screenshots have been rearranged for visualizing purposes, and the colored labels have been added for clarity

DESCRIPTION

Adrenal Mass



MedPix®

Demographics

38 y.o. female

Caption

Large, heterogeneously contrast-enhancing left supra-renal mass, without distinct evidence of

CASE

Adrenal Mass: NSCL cancer recurrence with adrenal met

History

Status post Chem/XRT for Stage IIIB NSCL cancer.

Exam

Vague abdominal tenderness to palpation over the left upper quadrant. Labs negative for urinary or hepatic source.

Findings

Heterogeneous soft tissue density with heterogeneous enhancement on IV contrast, in the region of the left supra-renal abdomen. Mild mass effect without obvious capsular invasion.

Case Diagnosis

Adrenal Mass: NSCL cancer recurrence with adrenal met

TOPIC

Adrenal Mass

Disease Discussion

This case presents the differential diagnosis for a large (5 cm) solitary adrenal mass in an asymptomatic patient. Broad

Fig. 2 Among the manually removed images there are presentation's slide (a), manually (b) and automatic annotated images from DICOM viewer (c)

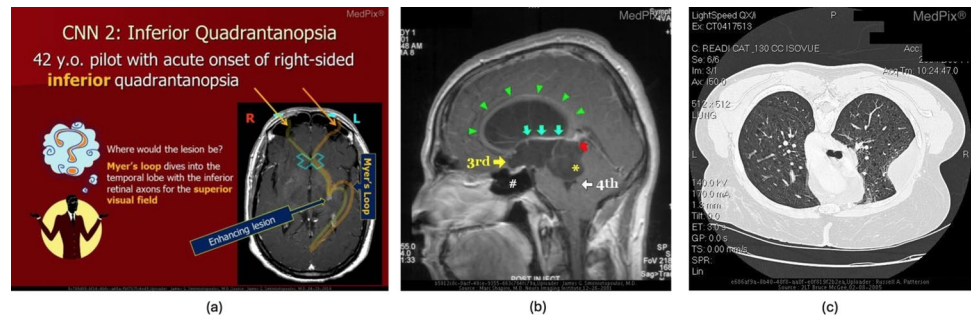


Fig. 3 Example of a JSON document in MedPix 2.0 as for description document

```
{
  "Type": "CT",
  "U_id": "MPX1009",
  "image": "MPX1009_synpic46283",
  "Description": {
    "Caption": "The prostate is enlarged with several calcifications noted within. No dominant prostate mass is evident.",
    "Plane": "Coronal",
    "Modality": "CT - noncontrast",
    "ACR Codes": "8.-1",
    "Sex": "male",
    "Age": "73"
  }
}
```

AI systems. Therefore we decided to create a brand new structured version of this data set because it represents a high quality source for AI-based medical applications. MedPix 2.0 has been built essentially as a MongoDB instance that is released along with a suitable GUI aimed both at general purpose querying and extracting training data for AI models. Referring to the Fig. 1, a MongoDB version of the data set was built using a semi-automated pipeline to create two kinds of JSON documents: the one collecting the information falling into the screenshots labeled as *DESCRIPTION*, and the one that gathers the information falling into the screenshots labeled both as *CASE* and *TOPIC*. As mentioned in the section 1, the choice of MongoDB is motivated by the aim of respecting the organisation of the main data, improving its readability. In fact, the pairs of images and texts are connected to each other by reference URIs, making it unnecessary to store the images within the documents as is typically done with a relational approach. In this way the creation of multimodal data sets is made more intuitive, facilitating any downstream application.

3.1 Data Set Extraction

We decided to focus on a part of MedPix[®] that involves cases from two diagnostic modalities, namely Computed Tomography (CT) and Magnetic Resonance Imaging (MRI). First

of all, the images in the considered split were downloaded via Open-i⁵: among them, there were some noisy images such as teaching materials or annotated images (Fig. 2) that are inconsistent as input image for training a VLM for which we expect a clear CT or MRI. In particular, this process of curing the data set by removing noisy samples, was performed via manual direct inspection of all the automatically downloaded samples. This was the only non-automated task performed during the data set extraction process; therefore, we refer to our approach as semi-automated. This selection was crucial, because the use of images with artefacts, such as the ones shown in the figure (e.g. symbols, annotations), would not allow a correct visual feature extraction, leading to relevant biases in the classification process of the multimodal model.

A subsequent automatic scraping pipeline was implemented to extract the textual data related to the selected images using Selenium⁶ and BeautifulSoup⁷. Finally, two kinds of JSON documents were devised to store, respectively, the information strictly connected with the images (*descriptions* document) and the one related to a

⁵ <https://openi.nlm.nih.gov/s>.

⁶ <https://www.selenium.dev/>.

⁷ <https://www.crummy.com/software/BeautifulSoup/bs4/doc/>.

Fig. 4 Example of a JSON document in MedPix 2.0 as for *case-topic document*

```
{
  "U_id": "MPX1009",
  "TAC": ["MPX1009_synpic46283", "MPX1009_synpic46295"],
  "MRI": [],
  "Case": {
    "Title": "Bladder Diverticulum",
    "History": "73-year-old male with hematuria and numerous white blood cells found on UA",
    "Exam": "N/A",
    "Findings": "Bladder with thickened wall and diverticulum on the right. Diverticulum is mostly likely secondary to chronic outflow obstruction. Prostate enlargement.",
    "Differential Diagnosis": "Bladder Diverticulum",
    "Case Diagnosis": "Bladder Diverticulum",
    "Diagnosis By": "N/A"
  },
  "Topic": {
    "Title": "Bladder Diverticulum",
    "Disease Discussion": "Bladder diverticula most often occur as a result of outlet obstruction. Occasionally, a congenital weakness in the bladder wall adjacent to the ureteral orifice results in a diverticulum. This is termed a \"Hutch\" diverticulum. In children, outlet obstruction causing a diverticulum is rare and can be seen with urethral [...] .",
    "ACR Code": "8.9",
    "Category": "Diverticulum",
  }
}
```

clinical case (*case-topic document*). A one-to-many relation has been created between a clinical cases and images by embedding the `U_id` defined for a *case-topic document* in each `descriptions` document attached to each image related to the clinical case itself. An example of the two kinds of JSON documents is shown in Figs. 3 and 4.

3.2 MongoDB Representation

In order to process properly the data in MedPix 2.0, we built a MongoDB database to host all the JSON documents along with the images. The architectural choice stems not only from the nature of the data in MedPix 2.0 but also from considerations related to its high flexibility and scalability to distributed environments where also private multimodal medical data could be added to the original collections with the constraint of not being moved away from their generation site as it is the case of hospital generated information.

We built a MongoDB instance made by two collections, namely `Image_Descriptions`, containing the `descriptions` documents, and `Clinical_reports` which contain all the *case-topic* documents. In our

implementation, the images are stored in a separate folder, and they are accessed using a proper `file://` URL built starting from their `U_id`. Finally, a view called `Image_Reports` allows a direct access to both collections via their `U_id`.

A user-friendly GUI was built using PyQt5⁸ to allow an easy access to the database, querying it to obtain the desired data for visualization and/or download. As shown in Fig. 5a, it is possible to select either the collection or the view to be queried and, add the query input in the relative fields. In Fig. 5b, an example of the answer to a query is reported: samples matching the query are reported as a list of JSON objects and the user can save it or view a specific clinical case and/or image selected in list of the query results. An example of this view is shown in Fig. 6, where a MedPix[®] like GUI is reproduced to enforce usability for the users of the original website. In our GUI, the curated images scraped along with the texts are showed by default, but the user can

⁸ <https://www.riverbankcomputing.com/software/pyqt/>.

Medical Database

☒ Image_Descriptions
☐ Clinical_Reports
☐ Image_Reports

Key Fields

☒ Type

☐ U_id

☐ Image

☐ Location

☐ Location Category

☒ Description

Query Inputs

Insert Type:

CT

Insert Description.ACR Codes:

Insert Description.Age:

Insert Description.Caption:

Insert Description.Figure Part:

Insert Description.Modality:

Insert Description.Plane:

Insert Description.Sex:

female

Query
Download
View Image

```

('Type': 'CT', 'U_id': 'MPX1024', 'Image': 'MPX1024_synpic40275', 'Description': ('ACR Codes': '1-1', 'Age': '60', 'Caption': 'CT of the chest reveals an obstructing mass and resultant LUL collapse.', 'Figure Part': None, 'Modality': 'CT - noncontrast', 'Plane': 'Axial', 'Sex': 'female'), 'Location': 'Chest, Pulmonary', 'Location Category': 'Thorax')
('Type': 'CT', 'U_id': 'MPX1024', 'Image': 'MPX1024_synpic40277', 'Description': ('ACR Codes': '1-1', 'Age': '60', 'Caption': 'CT of the chest reveals an obstructing mass and resultant LUL collapse.', 'Figure Part': None, 'Modality': 'CT - noncontrast', 'Plane': 'Axial', 'Sex': 'female'), 'Location': 'Chest, Pulmonary', 'Location Category': 'Thorax')
('Type': 'CT', 'U_id': 'MPX1012', 'Image': 'MPX1012_synpic22949', 'Description': ('ACR Codes': '8.2', 'Age': '14', 'Caption': 'pelvis', 'Figure Part': None, 'Modality': 'CT - GI & IV Contrast', 'Plane': 'Axial', 'Sex': 'female'), 'Location': 'Genitourinary', 'Location Category': 'Reproductive and Urinary System')
('Type': 'CT', 'U_id': 'MPX1012', 'Image': 'MPX1012_synpic22950', 'Description': ('ACR Codes': '8.2', 'Age': '14', 'Caption': 'pelvis', 'Figure Part': None, 'Modality': 'CT - GI & IV Contrast', 'Plane': 'Axial', 'Sex': 'female'), 'Location': 'Genitourinary', 'Location Category': 'Reproductive and Urinary System')
('Type': 'CT', 'U_id': 'MPX1016', 'Image': 'MPX1016_synpic34317', 'Description': ('ACR Codes': '6.3', 'Age': '43', 'Caption': 'Contrast enhanced chest CT shows diffuse increased interstitial markings involving the right middle and lower lobes.', 'Figure Part': '2a', 'Modality': 'CT w/contrast (IV)', 'Plane': 'Axial', 'Sex': 'female'), 'Location': 'Chest, Pulmonary', 'Location Category': 'Thorax')
('Type': 'CT', 'U_id': 'MPX1016', 'Image': 'MPX1016_synpic34318', 'Description': ('ACR Codes': '6.3', 'Age': '43', 'Caption': 'Contrast enhanced chest CT shows pleural thickening/scarring of the posterior right lower lobe.', 'Figure Part': '2b', 'Modality': 'CT w/contrast (IV)', 'Plane': 'Axial', 'Sex': 'female'), 'Location': 'Chest, Pulmonary', 'Location Category': 'Thorax')
('Type': 'CT', 'U_id': 'MPX1031', 'Image': 'MPX1031_synpic20041', 'Description': ('ACR Codes': '6.2', 'Age': '74', 'Caption': 'CT (3mm collimation) shows peripheral interlobular septal thickening with scattered focal consolidation and 660.', 'Figure Part': '2', 'Modality': 'CT - noncontrast', 'Plane': 'Axial', 'Sex': 'female'), 'Location': 'Chest, Pulmonary', 'Location Category': 'Thorax')
('Type': 'CT', 'U_id': 'MPX1038', 'Image': 'MPX1038_synpic20727', 'Description': ('ACR Codes': '3.0', 'Age': '27', 'Caption': 'Selected axial CT images in bone window.', 'Figure Part': '1', 'Modality': 'CT - noncontrast', 'Plane': 'Axial', 'Sex': 'female'), 'Location': 'Brain and Neuro', 'Location Category': 'Head')
('Type': 'CT', 'U_id': 'MPX1038', 'Image': 'MPX1038_synpic20729', 'Description': ('ACR Codes': '3.0', 'Age': '27', 'Caption': 'Selected axial CT images in bone window.', 'Figure Part': '3', 'Modality': 'CT - noncontrast', 'Plane': 'Axial', 'Sex': 'female'), 'Location': 'Brain and Neuro', 'Location Category': 'Head')
('Type': 'CT', 'U_id': 'MPX1038', 'Image': 'MPX1038_synpic20730', 'Description': ('ACR Codes': '3.0', 'Age': '27', 'Caption': 'Selected axial CT images in bone window.', 'Figure Part': '4', 'Modality': 'CT - noncontrast', 'Plane': 'Axial', 'Sex': 'female'), 'Location': 'Brain and Neuro', 'Location Category': 'Head')
('Type': 'CT', 'U_id': 'MPX1038', 'Image': 'MPX1038_synpic20731', 'Description': ('ACR Codes': '3.0', 'Age': '27', 'Caption': 'Selected axial CT images in bone window.', 'Figure Part': '5', 'Modality': 'CT - noncontrast', 'Plane': 'Axial', 'Sex': 'female'), 'Location': 'Brain and Neuro', 'Location Category': 'Head')
('Type': 'CT', 'U_id': 'MPX1056', 'Image': 'MPX1056_synpic41083', 'Description': ('ACR Codes': '7.5', 'Age': '51', 'Caption': 'Large, diffusely infiltrated fatty liver with accessory left lobe. Compare to density of spleen. Multiple areas of focal sparing in left lobe that appears nodular.', 'Figure Part': None, 'Modality': 'CT w/contrast (IV)', 'Plane': 'Axial', 'Sex': 'female'), 'Location': 'Abdomen', 'Location Category': 'Abdomen')
('Type': 'CT', 'U_id': 'MPX1056', 'Image': 'MPX1056_synpic41084', 'Description': ('ACR Codes': '7.5', 'Age': '51', 'Caption': 'Large, diffusely infiltrated fatty liver with accessory left lobe. Compare to density of spleen. Multiple areas of focal sparing in left lobe that appears nodular.', 'Figure Part': None, 'Modality': 'CT w/contrast (IV)', 'Plane': 'Axial', 'Sex': 'female'), 'Location': 'Abdomen', 'Location Category': 'Abdomen')
('Type': 'CT', 'U_id': 'MPX1033', 'Image': 'MPX1033_synpic45450', 'Description': ('ACR Codes': '6.2', 'Age': '6', 'Caption': 'IM PROGRESS', 'Figure Part': None, 'Modality': 'CT - noncontrast', 'Plane': 'Axial', 'Sex': 'female'), 'Location': 'Chest, Pulmonary', 'Location Category': 'Thorax')
('Type': 'CT', 'U_id': 'MPX1033', 'Image': 'MPX1033_synpic45451', 'Description': ('ACR Codes': '6.2', 'Age': '6', 'Caption': 'IM PROGRESS', 'Figure Part': None, 'Modality': 'CT - noncontrast', 'Plane': 'Axial', 'Sex': 'female'), 'Location': 'Chest, Pulmonary', 'Location Category': 'Thorax')

```

Save to JSON

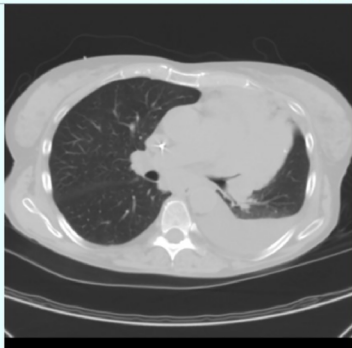
MPX1024_synpic40275
View

(a)
(b)

Fig. 5 An example of the interface for querying the database where the user asks for CT scans of female patients (a) and the provided output (b)

Fullscreen
Back
Next

Save Description JSON
Save Case/Topic JSON



Description

Type	CT
U_id	MPX1024
Image	MPX1024_synpic40275
Age	60
Caption	CT of the chest reveals an obstructing mass and resultant LUL collapse.
Sex	female
Location	Chest, Pulmonary
Location Category	Thorax

Case

Title Left upper lobe collapse caused by an enlarging, obstructing small cell lung carcinoma.

History 60-year-old woman presents with chest pain and shortness of breath.

Findings PA chest radiograph demonstrates left lung volume loss, silhouetting of the left cardiac border, and Luftsichel sign.[...]

Differential Diagnosis This combination of radiographic findings are consistent with LUL collapse and highly suspicious for an underlying endobronchial mass causing obstruction of the LUL bronchus.

Case diagnosis Left upper lobe collapse caused by an enlarging, obstructing small cell lung carcinoma.

Diagnosis by Lung biopsy proven Small Cell Lung Cancer.

Topic

Title Lung, lobar collapse (left upper lobe)

Disease Discussion Total left upper lobe atelectasis is caused by obstruction of the left upper lobe bronchus. This may be due to pulmonary mass, mucous plugging or foreign bodies. In older individuals, a pulmonary mass is a common cause PA films of left upper lobe atelectasis demonstrate opacification of the left upper lung field with obliteration of the left cardiac border. In addition, there is often a relative lucency adjacent to the aortic knob, which represents an overinflated lower lobe. This lucency is referred to as the Luftsichel sign. The lateral film demonstrates anterior displacement of the major fissure with increased anterior opacification. This patient had a previously known left upper lobe mass with total left upper lobe atelectasis proven by CT scan.

ACR Code 64.749

Category Obstruction or Stenosis

Fig. 6 An example of the view of a selected clinical case retrieved by the query posed in Fig. 5. The image, and the relative description are reported along with the information of the clinical case the image belongs to

choose to download the original (non curated) data related to the clinical case under investigation.

MedPix 2.0 and its querying interface can be a valid tool for both physicians and AI researchers since the desired data

Table 2 Summary of images and clinical cases for each split

Data set	#CT	#MR	Clinical case
Train	878	775	535
Validation	84	113	67
Test	100	100	69
Total	1062	988	671

can be easily downloaded for further application, like training Deep Learning (DL) models where a large amount of structured data are required. The structured output downloaded from the interface, can be used in turn to fill a fixed template, and to provide a VLM with a rich textual prompt. Also multimodal tasks can be addressed if the textual descriptions are combined with the corresponding image, as we will demonstrate in Sect. 4. Our re-arranged version of MedPix[®], contains the same information of the original data set, but allows a easier usage for DL application in text-only, image-only and multimodal setup. This flexible structure is also suitable for adding new samples that meet the characteristics of the data set, i.e. they should be a clear CT or MR scan with the associated textual data as in Figs. 3 and 4. The interface can be locally executed as reported in <https://github.com/CHILab1/MedPix-2.0/tree/main/MongoDB-UI>.

4 Training a VLM Using MedPix 2.0

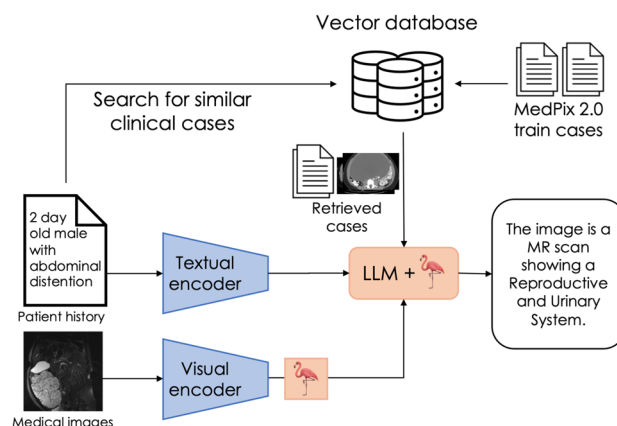
This section reports an application scenario for using MedPix 2.0. Specifically, a training pipeline for a multimodal model was implemented and, starting from the images, through a RAG-based approach, textual and visual embedding are generated to predict the scan modality and the body part shown in the input image. The output at this stage is furthermore used to query a generative Large Language Model (LLM) which will generate a consistent answer for the end user through the use of a Knowledge Graph (KG), the answer contains a diagnostic suggestion, leveraging the acquired knowledge on both scan modality and body part and the knowledge in the KG.

4.1 Data Preparation

The training phase of the model was conducted using MedPix 2.0, carefully divided into three splits: Train, Validation, and Test. In order to reduce data leakage, when splitting the data set into Train, Validation and Test splits, all the images referring to the same clinical case are reported in the same split. In addition, we ensured a correct stratification of the data as for balancing both the typology of the scanning modality, CT and MR scans, and the body part shown. Body parts were grouped into five macro categories,

Table 3 Summary of images divided per macro area of the body for each split

Data set	Abdomen	Head	RUS	Thorax	SaM
Train	264	742	127	263	257
Validation	23	66	20	30	58
Test	32	76	11	41	40
Total	319	884	158	334	355

**Fig. 7** Overview of DR-Minerva architecture

namely Abdomen, Head, Reproductive and Urinary System (RUS), Thorax and Spine and Muscles (SaM). Splits were created using a 80:20 in which Training and Testing split were derived; from a subsequent 50:50 split over the testing split, Validation and Test splits were expunged.

The data obtained from this division are shown in Tables 2 and 3.

4.2 RAG-Based Flamingo

Figure 7 illustrates the overall RAG architecture used to test MedPix 2.0, which is called DR-Minerva [5]. DR-Minerva relies on Flamingo [7] and Minerva [6], a new LLM trained from scratch on English and Italian data as part of the activities in the PNRR FAIR Transversal Project 2: “Vision, Language, and Multimodal Challenges”.⁹ FAIR TP2 is a collaborative research carried out by more than twenty Italian universities. The authors are engaged in developing VLM for the biomedical domain. We chose Minerva because it is totally open. Not only the weights, but also the architecture is free along with the training data. Furthermore, its training

⁹ <https://fondazione-fair.it/en/transversal-projects/tp2-vision-language-and-multimodal-challenges/>.

Table 4 The structure of the prompt is reported as well as the template of the corresponding expected answers

Prompt	Response
Given the following medical images and the patient history, provide information about the scanning modality and the body part shown in the image. $\langle \text{image} \rangle$ $\langle \text{age} \rangle \langle \text{sex} \rangle$ patient. $\langle \text{clinical history of the patient} \rangle$. [... context...]The image is a $\langle \text{scan type} \rangle$ scan showing a $\langle \text{body part} \rangle$. [... context...] $\langle \text{image} \rangle \langle \text{age} \rangle \langle \text{sex} \rangle$ patient. $\langle \text{clinical history of the patient} \rangle$	The image is a $\langle \text{predicted scan type} \rangle$ scan showing a $\langle \text{predicted body part} \rangle$

set is half English and half Italian, making it suitable for future experimentation using Italian data directly.

Flamingo was chosen as a suitable VLM since it has strong in-context learning capabilities that make it suitable to various domains [7]. Both in training and inference phase, visual tokens are incorporated with textual encodings, via a suitable trained cross-attention layer. More in detail, Flamingo allows the flexibility to adopt a custom visual and textual encoder, thus SOTA models can be easily used in this hybrid pipeline. On this basis, this approach was preferred to other VLMs, such as CLIP [30], which leverages an encoder-only models as for textual encoding, thus resulting in poor generative capabilities with respect to decoder-only models. In addition, the proposed classification via CLIP, is limited to pre-defined categories, which prevent the generative capabilities of the model in providing verbose answers. In our implementation, we used Open Flamingo [31] in the 3B parameter version,¹⁰ where CLIP ViT-L/14 [30] was employed as visual encoder and Minerva-3B¹¹ as language encoder. The Table 4 shows the prompt used to query the model in order to obtain the desired response, which consists of both the scanning modality and the body part shown in the input image.

We developed a suitable RAG component for DR-Minerva [8] to improve the overall performance of the model at inference time, due to a prompt enriched with relevant information for the query. The RAG technique allows you to construct a prompt that is based not only on the instructions and input provided, but also on relevant *ad hoc* information retrieved from a suitable Knowledge Base and is used as state of the art to respond to tasks of this type [32–34]. On this basis, the selected LLM can produce a coherent answer, thus mitigating any hallucinations due to the intrinsic knowledge in the LLM [8]. In addition, updated knowledge can be injected to the model via an in-context strategy [35], without requiring an additional fine-tuning of the model after a knowledge update [36]. Various strategies have been proposed to effectively enhance retriever capabilities, leveraging vector databases [9, 37]. Obviously, increasing retrieve

performances leads to better generated outputs and in optimizing computational costs.

Because AI models must be as precise as possible, especially in the medical sector, we queried DR-Minerva with both the target medical image and a template made up of some personal information about the patient (e.g. age and gender), followed by the patient's history. The most relevant clinical cases are then retrieved and added to the prompt in the form of few-shot learning examples, based on the patient's history. The RAG component selects the k-closest clinical cases to the query, more specifically, in our experiments k is equal to four. The value of this threshold was empirically found since it resulted in the optimal trade-off both for providing a sufficient context to the model and as for reaching satisfactory results. The vector database in the RAG component has been filled suitably with textual information from clinical cases, which belong to the train split, and involve the age and gender of the patient, the relative clinical history, and the doctor's diagnosis.

Our RAG is built using the LangChain framework,¹² which in turn employs a FAISS vector database [38], where the data are saved using Linq-Embed-Mistral [39]. This model is regarded as the best in the Massive Text Embedding Benchmark (MTEB) [40] for Information Retrieval.¹³

Both text-encoding and text generation are demanded to Minerva LLM, a transformer decoder-only LLM ([41]) built upon the Mistral architecture ([42]) which was trained from scratch on 660B tokens, equally balanced between English and Italian. Flamingo leverages a perceiver-resampler network to extract visual tokens from the visual encoding. In turn, visual tokens are employed to condition the textual output generation by inserting cross-attention layers between the existing pretrained ones and frozen LLM layers [7].

4.3 Knowledge Graph Definition

MedPix 2.0 is actually a huge archive of historical clinical cases that can be exploited as external knowledge for sophisticated textual predictions. In addition to the information

¹⁰ <https://huggingface.co/openflamingo/OpenFlamingo-3B-vitl-mpt1b>.

¹¹ <https://huggingface.co/sapienzanlp/Minerva-3B-base-v1.0>.

¹² <https://www.langchain.com/>.

¹³ As in <https://huggingface.co/spaces/mteb/leaderboard> in June 2024.

related to the particular case reported in the *Case* section, such as *Findings* and *Discussion*, there is also more general information related to the disease, like *Disease Discussion* in the *Topic* section (Figs. 3, 4)). Due to the inherent domain-specific semantics of this kind of textual information, classical RAG relying on vector databases is not efficient for precise retrieval. As a consequence, we decided to build a KG, which relies on both case-specific and general medical information that can be retrieved at inference time to generate a diagnostic suggestion for the provided input. The whole process for creating the KG and inferring suggestions is depicted in Fig. 8

To allow efficient relation extraction for building the KG, clinical cases belonging to the training set were arranged as documents according to the following template:

```
<U_id> is a clinical report of a <age><sex> patient suffering from a
<disease> displayed in <scan modality>.
<clinical history of the patient>
The disease <disease name> located in <specific body part>
(<body part>).
<clinical history of the patient>
<Treatment and followup recommended from the doctor>.
About <disease> we can say that: <disease discussion>.
```

We used LlamaIndex [43] to generate the KG. In particular, the KG is built starting from document chunks whose fixed size is 4096 tokens. At most n triplets were extracted from each chunk to form the nodes and edges of the KG. Triplets were extracted using the default LlamaIndex query posed to Llama 3.1 8B Instruct [12].

Llama 3.1 8B Instruct was chosen since it is an open-source and multi-lingual model trained with instruction-tuning strategy. More in detail, instruction tuned models, have the ability to provide a coherent output given a well-defined instruction, optionally enriched via some examples [35, 44]. This peculiar characteristics, made it suitable for both KG creation and the subsequent answer generation: both tasks rely on the capabilities of an external LLM to output a coherent answer leveraging the instruction. In addition, no further costly fine-tuning is requested at this step. No other models

were considered at this step, since, at the time of developing the model: this was the latest multilingual instruct model available, trained with a Supervised Fine-Tuning Strategy with Direct Preference Optimization [45] for human preference alignment [12].

We created six different KGs according to the maximum number of triplets to be extracted from each chunk (3, 5, 10) and the option to process the chunk in lowercase or not. The hyperparameter setups are reported in Table 5 and an example figure of the generated subgraph is shown in Fig. 9.

The entire process of creation and inference over the KGs was run on a cluster with 2 NVIDIA A100 64 GB GPUs.

5 Results

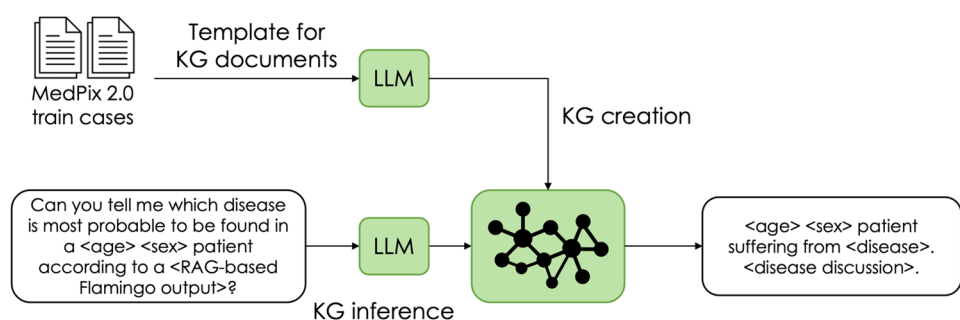
In this section are reported the results obtained with the developed multimodal pipeline, where an end-to-end VLM is used for medical diagnosis support. The architecture of this system is depicted in Fig. 10, and couples the original DR-Minerva model with the KG built as reported in Sect. 4.3. The main contribution is to assess the effective usability of the presented data set for complex tasks such as the proposed multimodal pipeline.

Performance evaluation was carried out by querying all the KGs with different hyperparameter choices, using prompts of increasing complexity. Specifically, three prompts were constructed, which were called *simple*,

Table 5 Overview of the hyperparameter setups used to generate the different versions of the KG

Setup name	Max rel. per chunk	Chunks in lower-case	# nodes	# edges
KG-s1	10	no	10,043	15,470
KG-s2	10	yes	10,706	16,179
KG-s3	3	no	5306	7385
KG-s4	3	yes	5172	6940
KG-s5	5	no	6473	9300
KG-s6	5	yes	6862	9521

Fig. 8 Overview of the overall process for KG creation and inference



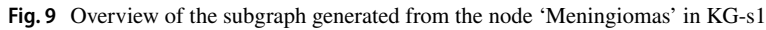
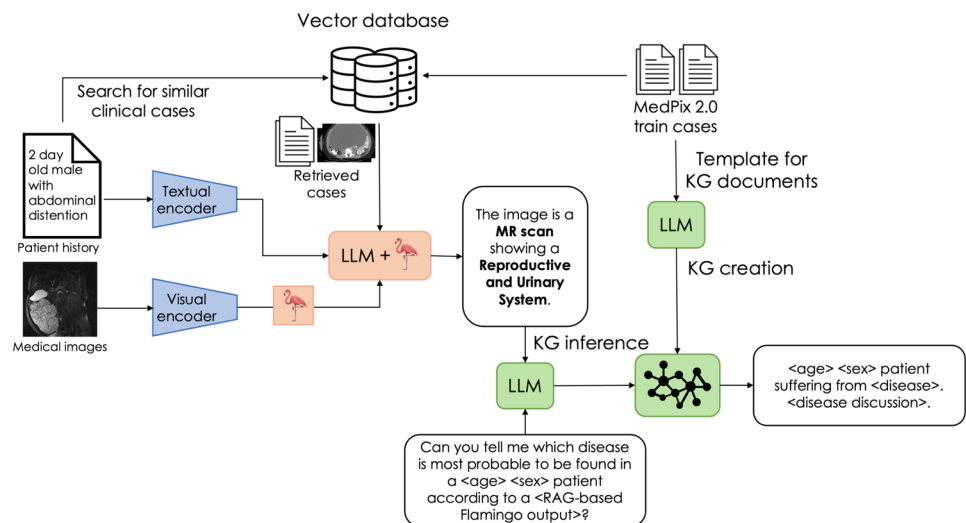


Fig. 10 Overview of the generated architecture with RAG-based Flamingo and KG creation and inference



mid-complex and *complex* respectively. All the prompts are reported as follows.

Prompt *simple*:

Can you tell me which diseases is most probable to be found in a patient having a {DR-Minerva output}?

Prompt *mid-complex*:

Can you tell me which diseases is most probable to be found in a $\langle \text{age} \rangle \langle \text{sex} \rangle$ patient according to a $\langle \text{DR-Minerva output} \rangle$?

Prompt *complex*:

Can you tell me which diseases are the most probable to be found in a ⟨age⟩ ⟨sex⟩ patient according to a ⟨DR-Minerva output⟩? Consider also the following additional information about the patient. ⟨history⟩.

The three prompts use the information provided by DR-Minerva, which consists in the scanning modality and the body part shown in the input image. Prompts were designed to provide more and more information to the KG, as their complexity increases. The maximum complexity will be reached in the last prompt, where we want to emulate the query that a doctor might pose to a diagnostic support system. In this case, information concerning the patient’s clinical *history* is used in addition to the information extrapolated from the images supplied to the multimodal system.

The availability of a test set extrapolated from MedPix 2.0 allowed us to generate a golden answer according the following template.

$\langle \text{age} \rangle \langle \text{sex} \rangle \text{patient suffering from } \langle \text{disease} \rangle. \langle \text{disease discussion} \rangle.$

Golden answers were used for the subsequent performance evaluation phase. Several metrics exist in the literature for automatic evaluation of LLM-generated text with respect to a set of reference or “golden” answers. These metrics score the degree of overlap between the reference answer $r = \{r_1, \dots, r_m\}$ and the hypothesis $h = \{h_1, \dots, h_n\}$ generated by the model. There are two approaches in this respect: character, word, or n-gram overlap and embedding similarity. BLEU [46] and ROUGE [47] belong to the first category. They are generally more tailored for evaluating Question-Answering and Machine Translation systems, for which a well-defined answer is known. Such measures require the generated text h to adhere strictly to the golden answer r . Despite the validity of these metrics, are not meaningful for evaluation purposes. In this context, a semantically comparison is required for the proposed task, with other metrics.

In the process of generating clinical text to support a doctor in the diagnostic process, the same meaning can be expressed in a variety of forms that depend on many factors. Some of these factors are the available knowledge about the particular case when the diagnosis is formulated, the different clinical practices in writing medical records, and so on. As a consequence, despite the generated golden answers and the data set splits, we cannot expect the model to generate text exactly as the golden answer: multiple correct answers (as for the meaning) can be proposed with different linguistic nuances.

For these reasons, metrics based on embedding similarity scores like METEOR [48] and BERT score [49], are preferred. In this case the generic r_j and h_i are embeddings and not words or n-grams. In particular, we decided to report in detail the BERT score in this paper and other metrics are reported in the supplementary material. BERT score is a neural framework to compare the embeddings of the tokens in r and h in terms of classical Precision, Recall, and F1 measure, which assess a compliant evaluation with the traditional metrics in AI field. In contrast to a traditional classification performance evaluation approach, Recall, Precision and BERT’s F1 aim to evaluate the degree of overlap of the generated response with respect to the target response. In our specific case, this golden label is represented by the entire topic document containing the clinical information. Consequently, the BERT metrics are well suited to evaluating a specific task of the model, which is to provide a verbose response to the doctor that is as close as possible to the correct clinical terminology. Equation 1 reports the BERTs score formulation.

$$\text{BERT-P} = \frac{1}{n} \sum_{i=1}^n \max_{r_j} \text{Sim}(h_i, r_j)$$

$$\text{BERT-R} = \frac{1}{m} \sum_{j=1}^m \max_{h_i} \text{Sim}(r_j, h_i) \quad (1)$$

$$\text{BERT-F1} = 2 \cdot \frac{\text{BERT-P} \cdot \text{BERT-R}}{\text{BERT-P} + \text{BERT-R}}.$$

$\text{Sim}(r_j, h_i)$ is the cosine similarity between the j -th reference token embedding and the i -th hypothesis token embedding, respectively. Table 6 shows the results obtained using all the three prompts. We report the results with the other metrics (BLEU, ROUGE, METEOR) in the supplementary material. Here we only report BERT score since it is the most significant.

The results show that the prompt *simple* achieves best performance for the models using KG-s1 and KG-s2, which are the KGs with the highest number of relations per chunk. The prompt *complex* performs best for the model using KG-s6, and it is close to the prompt *simple* for the models using KG-s3. No particular prompt prevails over the others in the models using KG-s4 and KG-s5.

Given the structure of the KGs, one can argue that the *simple* prompt does not constrain the model too much in the generation phase, and allows for the construction of verbose responses whose context is controlled by the high number of connected document chunks in the KG. On the other hand, when the KG connects less document chunks, the best performance of the prompt *complex* in terms of BERT-F1 depends on the increased generation precision induced by the insertion of the patient $\langle \text{history} \rangle$.

At inference time, Llama 3.1 8B Instruct is invoked for answer generation by retrieving relevant information from the KGs, leveraging the given question. We used a low temperature of 0.00001, thus the generative capabilities of the model are limited, and then, it is forced to produce an answer as close as possible to the retrieved information. Moreover, to prevent the model from generating loops, the `no_repeat_ngram_size` parameter was set to 2. The query strategy over the KG was done using the `tree_summarize` response mode and a value equal to 5 as for `similarity_top_k` parameter, that is the default parameters for the retriever. We report four end-to-end inferences using the better prompts and KGs, namely *simple* prompt and KG-s1 (Fig. 11) and *complex* prompt and KG-s6 (Fig. 12). Both the image, the input prompt to the model, the generated answer and the BERT-F1 score are reported.

Table 6 Summary of BERT metrics calculated on all KG setups for each of the proposed prompts

KG-setup	Location	Simple			Mid			Complex		
		BERT-Precision	BERT-Recall	BERT-F1	BERT-Precision	BERT-Recall	BERT-F1	BERT-Precision	BERT-Recall	BERT-F1
KG-s1	Head	0.8027	0.7840	0.7932	0.7870	0.7848	0.7857	0.7971	0.7858	0.7913
	Thorax	0.8003	0.7801	0.79	0.7853	0.7806	0.7827	0.7875	0.7822	0.7847
	SaM	0.7888	0.7932	0.791	0.7964	0.7851	0.7905	0.7846	0.7885	0.7864
	Abdomen	0.8008	0.7855	0.793	0.7859	0.7782	0.7818	0.7843	0.7817	0.7829
	RUS	0.8050	0.7886	0.7967	0.7774	0.7787	0.778	0.7884	0.7766	0.7823
KG-s2	Head	0.8090	0.7900	0.7994	0.7985	0.7863	0.7921	0.7888	0.7802	0.7843
	Thorax	0.8039	0.7851	0.7943	0.7981	0.7746	0.7856	0.7873	0.7821	0.7846
	SaM	0.7981	0.7950	0.7964	0.7888	0.7908	0.7898	0.7949	0.7860	0.7903
	Abdomen	0.7980	0.7845	0.7912	0.7882	0.7830	0.7853	0.7902	0.7863	0.7882
	RUS	0.7971	0.7850	0.791	0.8011	0.7817	0.7912	0.8063	0.7890	0.7975
KG-s3	Head	0.7990	0.7855	0.7921	0.7991	0.7825	0.7905	0.8019	0.7872	0.7944
	Thorax	0.7987	0.7840	0.7912	0.7885	0.7800	0.784	0.7858	0.7856	0.7855
	SaM	0.8058	0.7958	0.8007	0.7820	0.7906	0.7862	0.7993	0.7902	0.7947
	Abdomen	0.7960	0.7874	0.7916	0.7815	0.7742	0.7775	0.7844	0.7846	0.784
	RUS	0.8021	0.7799	0.7908	0.7972	0.7781	0.7874	0.7985	0.7878	0.7929
KG-s4	Head	0.7989	0.7799	0.7891	0.7915	0.7829	0.787	0.7952	0.7843	0.7895
	Thorax	0.8031	0.7784	0.7905	0.7869	0.7764	0.7815	0.7922	0.7822	0.7868
	SaM	0.8023	0.7934	0.7977	0.7870	0.7874	0.7871	0.8006	0.7878	0.7939
	Abdomen	0.7989	0.7769	0.7875	0.7975	0.7837	0.7905	0.7779	0.7765	0.7771
	RUS	0.7979	0.7819	0.7898	0.7753	0.7798	0.7774	0.7937	0.7819	0.7877
KG-s5	Head	0.7999	0.7858	0.7927	0.7960	0.7876	0.7917	0.7945	0.7814	0.7877
	Thorax	0.7960	0.7808	0.7882	0.7915	0.7769	0.784	0.7883	0.7827	0.7852
	SaM	0.8037	0.7953	0.7995	0.7888	0.7932	0.7908	0.8049	0.7925	0.7985
	Abdomen	0.8012	0.7772	0.7889	0.7840	0.7795	0.7816	0.8028	0.7820	0.7921
	RUS	0.7861	0.7860	0.786	0.7876	0.7871	0.7873	0.8025	0.7898	0.796
KG-s6	Head	0.7910	0.7864	0.7885	0.7923	0.7801	0.786	0.7931	0.7799	0.7862
	Thorax	0.7859	0.7812	0.7835	0.7720	0.7754	0.7735	0.7879	0.7808	0.7842
	SaM	0.7832	0.7893	0.7861	0.7957	0.7893	0.7924	0.8106	0.7918	0.8009
	Abdomen	0.7807	0.7816	0.7811	0.7958	0.7875	0.7916	0.7886	0.7801	0.7841
	RUS	0.7973	0.7826	0.7898	0.7849	0.7842	0.7845	0.7994	0.7886	0.7938

Bold values indicate the best score for each KG and body part respectively

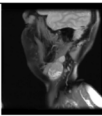
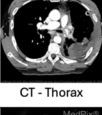
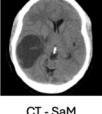
Image	Input prompt w/ age, sex of the patient	Generated answer w/ simple prompt over KG-s1	BERT-F1 score
 MR - Head	51 y.o. female patient.	Based on given information MP 142 and 230 are more related with Head and therefore more probably to find in CT of head. The most probaby disease in this case is MP-142 which is a Glioblastoma multiforme. [...]	0.779
 CT - Abdomen	52 y.o. male patient.	Based on provided context information and prior knowledge about Reporductive and Urinray system diseases, the most likely disease to appear on a MRI scan of the reproductive and urinary system is Urothelial carcinoma. Urotelial carcinoma is a type of cancer that affects the urinary bladder and is the most common type of bladder cancer in adults. It is more common in men than women and typically affects people over the age of 50. [...]	0.803
 CT - Thorax	38 y.o. male patient.	Based on the provided information, we cannot determine the most likely disease to find in patient with CT showing Thoracic region. However, based on general knowledge and the fact that Ewing Sarcoma and Ependyoma are more common in younger patients, it is possible that the patient might have Ewings Sarcomas or Epenymomas. But we need more information to determine which one is more probable.	0.780
 MR - RUS	1 y.o. male patient	The most likely disease to found is renal Cell carcinoma. Reasoning Skill: This question requires the ability to analyze the context and identify the most relevant information. In this case, the question is asking about the probability of finding a disease in an abdomen CT. The answer is based on the fact that the description of mrx1583 mentions that it's a CT w/contrast (IV) of an Abdonem and the location of renal carcinoma is in Genitounary (Reproductive And Urinary System). Therefore, it has a higher probability to have renal Carcinoma. [...]	0.801
 CT - SaM	52 y.o. female patient	Based on context, the disease most probable is Eprendyama. <nlp> Reason: Based on context information from multiple sources, Epenydoma is most commonly found in head and neck region. Therefore it can be concluded that the disease with highest probability of being present at a patient having ct scan showing Head will likely to have ependymoma. [...]	0.793

Fig. 11 One example for each body part of end-to-end queries to the entire system leveraging KG-s1 and simple prompt, along with inputs and the final generated answer

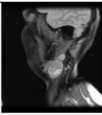
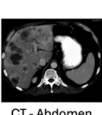
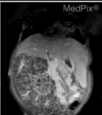
Image	Input prompt w/ age, sex and history of the patient	Generated answer w/ complex prompt over KG-s6	BERT-F1 score
 MR - Head	51 y.o. female patient. HPI: Presented with a 6 month hisotry of an asymptomatic left neck mass s/p failed course of abxtherapy and two non-diagnostic FNAs. PMH: Denies any h/o CA, XRT, chemo. Denies any h/o heart dz, lung dz, DM, SLE, scleroderma, [...]	The most probably disease is a benign neoplasm of the neck, specifically a lipomatous tumor. According to all information provided, a lymphadenopathy is less likely, because of two negative FNAB and failed abex therapy. A metastatic disease to neck from a known primary is very unlikely, due to lack of history or symptoms of a primary cancer site, and family cancer history is more suggestive of breast or colon cancer which are less common to metastasize to lymph nodes in the same side of primary. Finally, an inflammatory process is highly unlikely due the lack symptoms and negative history for infections. [...]	0.786
 CT - Abdomen	52 y.o. male patient. 52 yo male with RUQ pain and abnormal liver associated enzymes.	The most probabale disease that could be found in this male is Prostrat Cancer. Reason: 1- The age of patinet (52) makes him more likely for Prostrat Cancer 2- He presents symptoms such as difficulty starting or stopping urine flow which are common symptom related To prostata Cancer 3- He has RU pain which can be associated with liver and liver is affected by metastasized Prosta Cancer cells. [...]	0.800
 CT - Thorax	38 y.o. male patient. Thirty-eight year old male presents to primary care complaining of chronic cough and shortness of breath that has significantly progressed over several days	Based on the provided information, it is not possible to determine the most likely disease for a patient with symptoms of cough, short breath, chronic, progressive over days, 37 y. old, male, thorax CT. The information provided is too general and does not provide specific details about any disease that would be most relevant to this patient. However, based on general medical knowledge and the symptoms described, some possible diseases that could be considered include: Pneumonia, Pulmonary Embolism, Lung Cancer, Tuberculosis, etc. [...]	0.807
 MR - RUS	1 y.o. male patient 2 day old male with abdominal distention	Based on the provided information, it is not possible to determine the most likely disease to find in the 4 y. o. male patient. The information provided is about a different patient (0y.o) and a CT- non contrast scan of head. There is no information available about an abdominal CT-scan of this patient or any other patient with the same age and gender. However, based on general knowledge and the additional provided context, the diseases that are most common in children with an age range of around 6 months to 10 years are: Intussuseption, Appendicitis, and Gastrointestinal obstruction. [...]	0.803
 CT - SaM	52 y.o. female. 52 yo woman presented with headaches and mental status changes. Hx of L breast SCCa three years ago treated at that time with lumpectomy & chemoradiotherapy.	The most possible disease in this patient, based in the CT Scan and the clinical history, is a metastatic disease from breast Cancer. This conclusion is based in that the woman has a history of breast cancer and has symptoms that are consistent with metastases. Therefore, the final diagnosis is Metatasis.	0.820

Fig. 12 One example for each body part of end-to-end queries to the entire system leveraging KG-s6 and complex prompt, along with inputs and the final generated answer

6 Limitations

MedPix 2.0 is proposed as a multimodal data set for training deep neural architectures. To demonstrate the robustness and ease of implementing the data set, a model was proposed that combined a RAG-based Flamingo, Dr-Minerva, (created by the authors in a previous work) in combination with Llama 3.1 for the generation of verbose responses. As can be seen from the results obtained and evaluated in terms of BERT Precision, Recall and F1, the result is robust enough to allow an initial experimental version of a multimodal model to achieve good performance in generating a coherent verbose response. Despite the promising results, the construction of a reliable model is not the main topic of the proposed work, which is why validation by doctors has not been carried out, a step that is a fundamental in proposing a model to support diagnosis. Despite the simplicity with which MedPix 2.0 can be queried and used in its Image-only, Text-only and Multimodal versions, its integration with other data sets can improve its structure, making it more complete and maximizing its effectiveness in AI training tasks for diagnostic support.

7 Conclusions and Future Work

In this work we presented MedPix 2.0, a multimodal data set of clinical reports, CT and MR scans. We devised a semi-automated pipeline to download and curate the images in the original data sets, while structuring the textual information as a set of JSON document collections that had been used to build a proper MongoDB instance. The NoSQL version of the data set can be accessed and queried with a usable GUI that has been developed purposely. Using the GUI one can browse the data set in the same manner as in the original data set, and can download the structured output of the query that is suitable for training AI models. In fact, MedPix[®] and its MongoDB interface, represent in our view a relevant starting point for the development of AI multimodal models in the medical domain, such as Information Extraction systems tailored for clinical reports, automated analysis of the medical images, or Generative AI models for clinical report generation as part of a Medical Decision Support System. All these systems can rely on MedPix 2.0 as a structured source of data containing both clinical cases and medical explanations about the disease under investigation.

To demonstrate this point, we recalled the architecture of DR-Minerva, a RAG-based Flamingo architecture

already developed by the authors, and presented a brand-new VLM architecture that couples DR-Minerva with a KG created automatically relying on Llama 3.1 8B Instruct. Different KG structures were tested, starting from the MedPix 2.0 test set along with prompts of increasing complexity as regards the context provided to the model. The results we obtained are very satisfactory but we are actively working to improve the model's generation ability by means of a structural and semi-automatic organization of the KG, relying on external medical ontologies.

The scalability of the developed MongoDB database makes it suitable for future extensions with the possibility to add brand new clinical cases or existing cases from data set mentioned in Sect. 2, which have to be compliant with privacy regulations and follow the required information structure. Moreover, the inherent distributed nature of MongoDB allows for creating huge databases across different wards where the data owned by a single institution do not need to be explicitly moved out of the hospital thus violating privacy regulations. The structure of MedPix 2.0 could also serve as a guide to develop suitable connectors to share allowed data in the EHDS.

New cases can also be easily added to the KG, thus expanding the clinical knowledge base used by the system in the inference phase. Further improvement is being done on the GUI, providing the user with advanced data visualization tools like the possibility to interactively compare similar cases, thus helping physicians during the diagnostic phase.

Supplementary information Tables with the results obtained from the experiments for the clinical cases for the three prompts and for all KG setups are given in the supplementary file.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s41019-025-00297-8>.

Acknowledgements This work is supported by the project B73C22000810001, project code ECS_00000022, “SAMOTHRACE” (Sicilian MicronanoTech Research And Innovation Center). Models are built on the Leonardo supercomputer with the support of CINECA-Italian Super Computing Resource Allocation, class C project IsrC_DOC-VLM (HP10C64J82).

Author Contributions All authors contributed to the study conception and design. Material preparation and data collection were performed by M.L.C., R.A. Experiments were performed by I.S. and S.C., they also cured the first draft of the manuscript. Scientific supervision and review was cured by R.P.

Funding This research was funded by “SAMOTHRACE” (Sicilian MicronanoTech Research And Innovation Center), project B73C22000810001, project code ECS 00000022.

Data and Code Availability Source code and is freely available at <https://github.com/CHILab1/MedPix-2.0.git>. The data used for the test and train can be downloaded free of charge from the zenodo repository at the following link <https://zenodo.org/records/12624810>.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Penedo AC (2024) The regulation of data spaces under the EU data strategy: towards the 'actification' of the fifth European freedom for data? *Eur J Law Technol* 15(1)
- Terzis P, Echeverria EOS (2023) Interoperability and governance in the European Health Data Space regulation. *Med Law Int*. <https://doi.org/10.1177/09685332231165692>
- El Emam K, Rodgers S, Malin B (2015) Anonymising and sharing individual patient data. *The BMJ* 350:1139. <https://doi.org/10.1136/bmj.h1139>
- Chevrier R, Foufi V, Gaudet-Blavignac C, Robert A, Lovis C (2019) Use and understanding of anonymization and de-identification in the biomedical literature: scoping review. *J Med Intern Res* 21(5):13484. <https://doi.org/10.2196/13484>
- Siragusa I, Contino S, Pirrone R (2025) DR-Minerva: a multimodal language model based on Minerva for diagnostic information retrieval. In: *AIxIA 2024—advances in artificial intelligence*. Springer, Cham, pp 288–300
- Orlando R, Moroni L, Cabot P-LH, Barba E, Conia S, Orlandini S, Fiameni G, Navigli R (2024) Minerva LLMs: the first family of large language models trained from scratch on Italian data. In: *Proceedings of the 10th Italian conference on computational linguistics (CLiC-it 2024)*
- Alayrac J-B, Donahue J, Luc P, Miech A, Barr I, Hasson Y, Lenc K, Mensch A, Millican K, Reynolds M, Ring R, Rutherford E, Cabi S, Han T, Gong Z, Samangooei S, Monteiro M, Menick J, Borgeaud S, Brock A, Nematzadeh A, Sharifzadeh S, Binkowski M, Barreira R, Vinyals O, Zisserman A, Simonyan K (2022) Flamingo: a visual language model for few-shot learning
- Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, Goyal N, Küttler H, Lewis M, Yih W-T, Rocktäschel T et al (2020) Retrieval-augmented generation for knowledge-intensive NLP tasks. *Adv Neural Inf Process Syst*
- Wang C, Ong J, Wang C, Ong H, Cheng R, Ong D (2024) Potential for GPT technology to optimize future clinical decision-making using retrieval-augmented generation. *Ann Biomed Eng*
- Ghanbari Haez S, Segala M, Bellan P, Magnolini S, Sanna L, Consolandi M, Dragoni M (2024) A retrieval-augmented generation strategy to enhance medical chatbot reliability. In: *Artificial intelligence in medicine*
- Wu J, Zhu J, Qi Y, Chen J, Xu M, Menolascina F, Grau V (2024) Medical graph RAG: towards safe medical large language model via graph retrieval-augmented generation. [arxiv:2408.04187](https://arxiv.org/abs/2408.04187)
- Llama Team A.M (2024) The Llama 3 herd of models. [arXiv:2407.21783](https://arxiv.org/abs/2407.21783)
- Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014) Microsoft coco: common objects in context. In: *Computer vision—ECCV 2014: 13th European conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V* 13. Springer, pp 740–755
- Pelka O, Koitka S, Rückert J, Nensa F, Friedrich C.M (2018) Radiology objects in context (ROCO): a multimodal image dataset. In: *Intravascular imaging and computer assisted stenting and large-scale annotation of biomedical data and expert label synthesis: 7th joint international workshop, CVII-STENT 2018 and third international workshop, LABELS 2018, held in conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings* 3. Springer, pp 180–189
- Subramanian S, Wang LL, Mehta S, Bogin B, Zuylen M, Parasa S, Singh S, Gardner M, Hajishirzi H (2020) Medcat: a dataset of medical images, captions, and textual references. *arXiv preprint arXiv:2010.06000*
- Lin W, Zhao Z, Zhang X, Wu C, Zhang Y, Wang Y, Xie W (2023) PMC-clip: contrastive language-image pre-training using biomedical documents. In: *International conference on medical image computing and computer-assisted intervention*. Springer, pp 525–536
- Lau JJ, Gayen S, Ben Abacha A, Demner-Fushman D (2018) A dataset of clinically generated visual questions and answers about radiology images. *Sci Data* 5(1):1–10
- He X, Zhang Y, Mou L, Xing E, Xie P (2020) Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*
- Johnson AE, Pollard TJ, Berkowitz SJ, Greenbaum NR, Lungren MP, Deng C-Y, Mark RG, Horng S (2019) MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci Data* 6(1):317
- Demner-Fushman D, Kohli MD, Rosenman MB, Shooshan SE, Rodriguez L, Antani S, Thoma GR, McDonald CJ (2016) Preparing a collection of radiology examinations for distribution and retrieval. *J Am Med Inf Assoc* 23(2):304–310
- Liu B, Zhan L-M, Xu L, Ma L, Yang Y, Wu X-M (2021) Slake: a semantically-labeled knowledge-enhanced dataset for medical visual question answering. In: *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*. IEEE, pp 1650–1654
- Huang S-C, Huo Z, Steinberg E, Chiang C-C, Lungren MP, Langlotz CP, Yeung S, Shah NH, Fries JA (2023) INSPECT: a multimodal dataset for pulmonary embolism diagnosis and prognosis. *arXiv preprint arXiv:2311.10798*
- Chaudhry HAH, Renzulli R, Perlo D, Santinelli F, Tibaldi S, Cristiano C, Grosso M, Limerutti G, Fiandrotti A, Grangetto M et al (2022) Unitochest: a lung image dataset for segmentation of cancerous nodules on CT scans. In: *International conference on image analysis and processing*. Springer, pp 185–196
- ...Jack CR Jr, Bernstein MA, Fox NC, Thompson P, Alexander G, Harvey D, Borowski B, Britson PJ, Whitwell JL, Ward C, Dale AM, Felmlee JP, Gunter JL, Hill DLG, Killiany R, Schuff N, Fox-Bosetti S, Lin C, Studholme C, DeCarli CS, Krueger G, Ward HA, Metzger GJ, Scott KT, Mallozzi R, Blezek D, Levy J, Debbins JP, Fleisher AS, Albert M, Green R, Bartzokis G, Glover G, Mugler J, Weiner MW (2008) The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *J Magn Reson Imaging* 27(4):685–691. <https://doi.org/10.1002/jmri.21049>
- Halling-Brown MD, Warren LM, Ward D, Lewis E, Mackenzie A, Wallis MG, Wilkinson LS, Given-Wilson RM, McAviney R, Young KC (2021) OPTIMAM mammography image database: a large-scale resource of mammography images and clinical data. *Radiol Artif Intell* 3(1):200103. <https://doi.org/10.1148/ryai.2020200103>
- LaMontagne PJ, Benzinger TL, Morris JC, Keefe S, Hornbeck R, Xiong C, Grant E, Hassenstab J, Moulder K, Vlassenko AG, Raichle ME, Cruchaga C, Marcus D (2019) OASIS-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and Alzheimer disease. *medRxiv*. <https://doi.org/10.1101/2019.12.13.19014902>

27. Schulz S, Ševa J, Rodriguez S, Ostendorff M, Rehm G (2020) Named entities in medical case reports: corpus and experiments. In: Proceedings of the twelfth language resources and evaluation conference. European Language Resources Association, Marseille, France, pp 4495–4500. <https://aclanthology.org/2020.lrec-1.553>
28. Gava U, D'Agata F, Bennink E, Tartaglione E, Perlo D, Vernone A, Bertolino F, Ficiara E, Cicerale A, Pizzagalli F, Guiot C, Grangotto M, Bergui M. UniTOBrain. <https://doi.org/10.21227/x8ea-vh16>
29. Magnini B, Altuna B, Lavelli A, Speranza M, Zanoli R (2020) The E3C project: collection and annotation of a multilingual corpus of clinical cases. In: Proceedings of the seventh Italian conference on computational linguistics CLiC-it 2020
30. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G, Sutskever I (2021) Learning transferable visual models from natural language supervision
31. Awadalla A, Gao I, Gardner J, Hessel J, Hanafy Y, Zhu W, Marathe K, Bitton Y, Gadre S, Sagawa S, Jitsev J, Kornblith S, Koh PW, Ilharco G, Wortsman M, Schmidt L (2023) OpenFlamingo: an open-source framework for training large autoregressive vision-language models. arXiv preprint [arXiv:2308.01390](https://arxiv.org/abs/2308.01390)
32. Jiang X, Fang Y, Qiu R, Zhang H, Xu Y, Chen H, Zhang W, Zhang R, Fang Y, Chu X, Zhao J, Wang Y (2024) TC-RAG: turing-complete rag's case study on medical LLM systems. [arXiv:2408.09199](https://arxiv.org/abs/2408.09199). <https://doi.org/10.48550/arXiv.2408.09199>
33. Salemi A, Zamani H (2024) Comparing retrieval-augmentation and parameter-efficient fine-tuning for privacy-preserving personalization of large language models. [arXiv:2409.09510](https://arxiv.org/abs/2409.09510) [cs]. <https://doi.org/10.48550/arXiv.2409.09510>
34. Lee J, Ahn S, Kim D, Kim D (2024) Performance comparison of retrieval-augmented generation and fine-tuned large language models for construction safety management knowledge retrieval. *Autom Constr* 168:105846. <https://doi.org/10.1016/j.autcon.2024.105846>
35. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A et al (2020) Language models are few-shot learners. *Adv Neural Inf Process Syst*
36. Peng B, Galley M, He P, Cheng H, Xie Y, Hu Y, Huang Q, Liden L, Yu Z, Chen W, Gao J (2023) Check your facts and try again: improving large language models with external knowledge and automated feedback. [arXiv:2302.12813](https://arxiv.org/abs/2302.12813)
37. Vividelli S, Ramachandran M, Dharunbalaji A (2024) Efficiency-driven custom chatbot development: unleashing LangChain, RAG, and performance-optimized LLM fusion. *Mater Contin Comput*
38. Douze M, Guzhva A, Deng C, Johnson J, Szilvasy G, Mazaré P-E, Lomeli M, Hosseini L, Jégou H (2024) The Faiss library. [arXiv:2401.08281](https://arxiv.org/abs/2401.08281)
39. Junseong K, Seolhwa L, Jihoon K, Sangmo G, Yejin K, Minkyung C, Jy-yong S, Chanyeol C (2024) Linq-Embed-Mistral: Elevating text retrieval with improved GPT data through task-specific control and quality refinement. *LinQ AI Research Blog*. <https://getlinq.com/blog/linq-embed-mistral/>
40. Muennighoff N, Tazi N, Magne L, Reimers N (2022) MTEB: massive text embedding benchmark. arXiv preprint [arXiv:2210.07316](https://arxiv.org/abs/2210.07316). <https://doi.org/10.48550/ARXIV.2210.07316>
41. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. In: Proceedings of the 31st international conference on neural information processing systems. NIPS'17. Curran Associates Inc., Red Hook, NY, USA, pp 6000–6010
42. Jiang AQ, Sablayrolles A, Mensch A, Bamford C, Chaplot DS, Casas D, Bressand F, Lengyel G, Lample G, Saulnier L, Lavaud LR, Lachaux M-A, Stock P, Scao TL, Lavril T, Wang T, Lacroix T, Sayed WE (2023) Mistral 7B. [arXiv:2310.06825](https://arxiv.org/abs/2310.06825)
43. Liu J. LlamaIndex. <https://doi.org/10.5281/zenodo.1234>. https://github.com/jerryliu/llama_index
44. Mialon G, Dessì R, Lomeli M, Nalmpantis C, Pasunuru R, Raileanu R, Rozière B, Schick T, Dwivedi-Yu J, Celikyilmaz A, Grave E, LeCun Y, Scialom T (2023) Augmented language models: a survey. [arXiv:2302.07842](https://arxiv.org/abs/2302.07842)
45. Rafailov R, Sharma A, Mitchell E, Ermon S, Manning C.D, Finn C (2024) Direct preference optimization: your language model is secretly a reward model. [arXiv:2305.18290](https://arxiv.org/abs/2305.18290)
46. Papineni K, Roukos S, Ward T, Zhu W-J (2002) BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the association for computational linguistics, pp 311–318
47. Lin C-Y (2004) Rouge: a package for automatic evaluation of summaries. In: Text summarization branches out, pp 74–81
48. Banerjee S, Lavie A (2005) METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, pp 65–72
49. Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y (2020) BERTScore: evaluating text generation with BERT. In: International conference on learning representations. <https://openreview.net/forum?id=SkeHuCVFDr>