HeadCT-ONE: Enabling Granular and Controllable Automated Evaluation of Head CT Radiology Report Generation

Julián N. Acosta, MD* JULIAN_ACOSTA@HMS.HARVARD.EDU
Department of Biomedical Informatics, Harvard Medical School, USA

Xiaoman Zhang, PhD* XIAOMAN_ZHANG@HMS.HARVARD.EDU
Department of Biomedical Informatics, Harvard Medical School, USA

Siddhant Dogra, MD SIDDHANT.DOGRA@NYULANGONE.ORG
Department of Radiology, NYU Langone Health, USA

Hong-Yu Zhou, PhD HONGYU_ZHOU@HMS.HARVARD.EDU
Department of Biomedical Informatics, Harvard Medical School, USA

Seyedmehdi Payabvash, MD SP4479@CUMC.COLUMBIA.EDU
Department of Radiology, Columbia University, USA

Guido J. Falcone, MD, ScD, MPH GUIDO.FALCONE@YALE.EDU
Department of Neurology, Yale University School of Medicine, USA

Eric K. Oermann, MD ERIC.OERMANN@NYULANGONE.ORG
Department of Neurosurgery, NYU Langone Health, USA

Pranav Rajpurkar, PhD† PRANAV_RAJPURKAR@HMS.HARVARD.EDU
Department of Biomedical Informatics, Harvard Medical School, USA

Abstract

We present Head CT Ontology Normalized Evaluation (HeadCT-ONE), a metric for evaluating head CT report generation through ontology-normalized entity and relation extraction. HeadCT-ONE enhances current information extraction derived metrics (such as RadGraph F1) by implementing entity normalization through domain-specific ontologies, addressing radiological language variability. HeadCT-ONE compares normalized entities and relations, allowing for controllable weighting of different entity types or specific entities. Through experiments on head CT reports from three health systems, we show that HeadCT-ONE's normalization and weighting approach improves the capture of semantically equivalent reports, better distinguishes between normal and abnormal reports, and aligns with radiologists' assessment of clinically significant errors, while offering flexibility to prioritize specific aspects of report content. Our results demonstrate how HeadCT-ONE enables more flexible, controllable, and granular automated evaluation of head CT reports.

Data and Code Availability Data utilized in the study is from a proprietary de-identified medical imaging dataset from multiple institutions in the U.S. and will not be made available to other researchers. The code is available at https://github.com/rajpurkarlab/HeadCT-ONE.

Institutional Review Board (IRB) This research utilizes fully de-identified medical data and therefore does not require IRB approval.

1. Introduction

Automated radiology report generation, leveraging artificial intelligence (AI) to produce descriptive text from medical images, has gained significant attention due to its potential to streamline clinical workflows and improve patient care (Moor et al., 2023; Rajpurkar and Lungren, 2023; Sloan et al., 2024). As AI systems approach human-like performance in generating radiology reports from images, the development of robust, automated evaluation metrics be-

* These authors contributed equally
† Corresponding Author

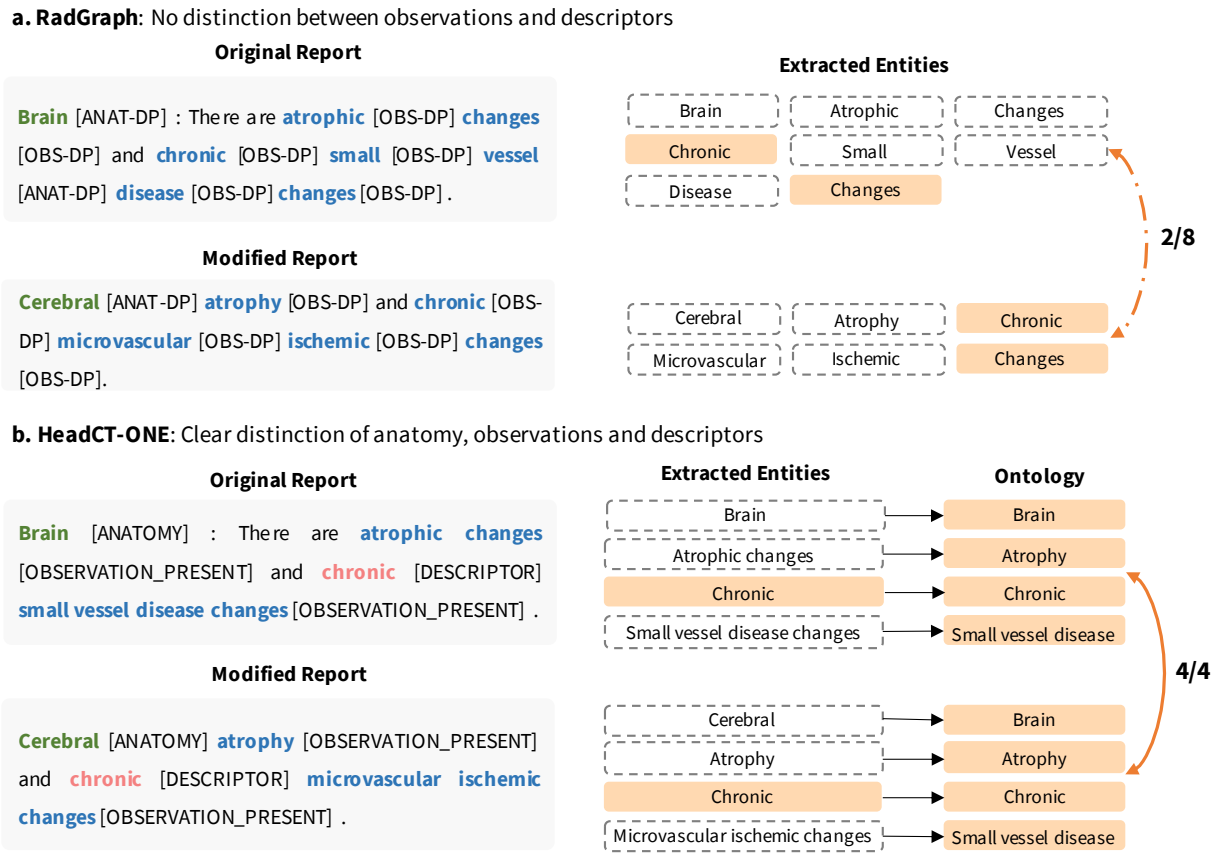

Figure 1: Comparison of entity extraction results between RadGraph F1 and HeadCT-ONE for two reports with different writing styles but identical findings. With ontology normalization, HeadCT-ONE's extractions are identical for both reports, while RadGraph F1 only matches 1/4 of the entities.

comes paramount. However, evaluating AI-generated radiology reports presents unique challenges due to the specialized nature of medical language and the semantic complexity of these reports.

Traditional natural language generation metrics, such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004), often fall short when applied to radiology, primarily focusing on lexical overlap and struggling to capture nuanced, semantic similarities crucial in medical reporting. One approach to address this limitation is the use of classification labels, such as the CheXpert F1 score (Irvin et al., 2019), which evaluates the presence or absence of specific clinical findings. However, while this method sidesteps lexical variation issues, it fails to capture the level of detailed crucial in radiology reports. More sophisticated approaches include RadGraph F1 (Yu et al., 2023), an evaluation metric based on RadGraph (Jain et al., 2021) that extracts entities and relations in reference and candidate reports and measure their overlap, providing a more meaningful evaluation but lacking robustness to variations in medical language (Figure 1), and embedding-based techniques (Zhang et al., 2020; Endo et al., 2021), which aim to capture semantic similarities by comparing vector representations of text, but output a single number, lacking explainability.

Large Language Model (LLM)-based evaluation methods have recently emerged as a promising approach (Huang et al., 2024; Chaves et al., 2024; Xie et al., 2024; Ostmeier et al., 2024), being more robust to stylistic differences (Banerjee et al., 2024) and aligning well with radiologists. However, current LLM-based metrics only ascertain error type and

their clinical significance, limiting the potential for more granular evaluation, for example, by weighting on different pathologies, anatomical locations or descriptors, which may be more relevant for different clinical scenarios or distinct models being evaluated.

We argue that automated information extraction, coupled with accurate normalization to domain-specific ontologies can address these limitations by providing a standardized framework for medical terminology, enabling fine-grained categorization of concepts, and allowing for customizable weighting of different entities. Further, this approach could still leverage the natural language understanding of LLMs at different steps (Gilbert et al., 2024), offering a more robust, explainable, and flexible approach to radiology report evaluation.

In this work:

1. We present a new information extraction framework that extends the RadGraph radiology report information extraction framework by introducing entity normalization, leveraging domain-specific ontologies for observations, anatomical terms, and descriptors. While this pipeline is focused on head CT, it is easily extendable to other CT modalities.

2. We introduce HeadCT-ONE (Ontology normalized evaluation), a radiology report generation evaluation metric measuring overlap between these normalized entities and relations in candidate and reference reports, allowing the metric to be robust to differences in radiological language. We demonstrate that HeadCT-ONE captures similarities between radiology reports more effectively, showing increased robustness to variations in style and language due to the normalization process, facilitating the distinction between normal and abnormal reports.

3. We introduce a weighting mechanism for different entity types or specific entities, allowing for the creation of metrics that focus on particular aspects of AI-generated reports. This controllability enables more nuanced evaluation tailored to specific clinical scenarios or model limitations. We illustrate how weighting on different entity types or specific entities allows our metric to focus on distinct issues in radiology report generation, and align better with radiologists' assessments of clinically significant errors.

Our work represents a significant step towards more flexible and aligned automated evaluations of radiology report generation, facilitating the development and clinical evaluation of AI systems for head CT radiology report generation.

2. Related Work

There have been multiple efforts to develop reliable metrics for radiology report generation. Traditional natural language generation metrics, such as BLEU, ROUGE, and METEOR primarily focus on lexical overlap and struggle to capture the nuanced, semantic similarities that are crucial in medical reporting. (Yu et al., 2023) introduced RadGraph F1, a more clinically meaningful evaluation metric based on extracting entities and relations from reports using RadGraph (Jain et al., 2021). However, the inherent variability in radiological language—where multiple phrases can express the same clinical finding—poses a challenge for this type of metric. This variability can lead to the underestimation of similarity between clinically equivalent but lexically different reports.

Embedding-based evaluation metrics, such as BERTScore (Zhang et al., 2020) and SemB score (Endo et al., 2021), have attempted to address these limitations by leveraging pre-trained language models to capture semantic similarities. While these approaches offer improvements over traditional lexical overlap metrics, they still face challenges in the medical domain. These metrics often lack explainability, making it difficult for clinicians to interpret and trust the scores. Additionally, when applied at the report level, they struggle to provide granular insights or allow for fine-grained control over the evaluation process, which is crucial in the nuanced field of radiology.

Large Language Model (LLM)-based evaluation approaches have emerged as a promising alternative for assessing radiology reports, with multiple methods recently introduced, including FineRadScore (Huang et al., 2024), CheXprompt (Chaves et al., 2024), DocLens (Xie et al., 2024) and GREEN (Ostmeier et al., 2024), offering several advantages over traditional metrics. These methods demonstrate improved alignment with human judgment and exhibit robustness to lexical variations, capturing semantic similarities that might elude conventional metrics. However, LLM-based evaluations come with their own set of challenges, being limited to the categorization of error types and their clinical significance, which they struggle to capture accurately.

Further, running these metrics at scale can be prohibitively expensive, especially when dealing with large datasets common in medical imaging. Moreover, the use of LLMs raises significant privacy concerns, particularly when handling sensitive medical data or proprietary datasets.

Zhao et al. (2024) introduced RaTEScore, demonstrating promise in utilizing a named entity recognition and synonym disambiguation approach, outperforming other metrics on its alignment with radiologists' scores. However, RaTEScore does not leverage ontologies nor defines a descriptor entity type, which limit the potential controllability of this score.

3. Method

3.1. Dataset

Primary dataset. We leverage a large proprietary dataset of 101,319 head CT and their corresponding radiology reports from three health systems in the United States and multiple sites within them, containing studies from 35,380 patients. Mean age at the time of study is 65.7 (SD 18.7), and 57% of the studies are from female patients. This dataset is private and will not be available to other researchers.

NER models training data. We randomly sampled 2,000 radiology reports from this dataset to train our NER models.

Experiments data. Original reports: We randomly sampled reports from the 20 sites with the most studies in the dataset until we obtained 1 normal and 1 abnormal report per site, resulting in 40 reports total. The normal/abnormal status of each report was determined by manual review. Additionally, we randomly sampled a non-overlapping group of 400 reports from these same 20 distinct sites.

Modified reports: We prompt GPT-4o (version gpt4o05132024) through the Azure OpenAI secure API to obtain 5 different versions of these reports: rephrased, any error, observation errors, anatomical errors, and descriptor errors. For rephrasing, the prompt instructs GPT-4o to rewrite each report using alternative wording and sentence structures while preserving its clinical content; for any error, the prompt instructs the model to introduce one or more arbitrary errors without specific constraints; for observation errors, the model is directed to introduce 1–3 errors by either omitting, adding, or negating abnormal observations; for anatomical errors, the model is instructed to alter anatomical details, for example by switching laterality or assigning an incorrect anatomical region; and for descriptor errors, the prompt directs the model to modify descriptive elements (e.g., size, shape, or density) without changing the core observation. Prompts are shown in the Appendix. A random sample of these modified reports was evaluated by a clinical expert to confirm that the introduced errors were appropriate.

3.2. Ontology

Our clinical team developed a specialized ontology for the three main entity types in our expanded information extraction schema. This development was led by a neurologist and a radiology trainee, involving iterative refining and final approval by a team of experienced clinicians, including a neuroradiologist, a neurologist, and a neurosurgeon.

Observation ontology: For the observation ontology, we began with an existing comprehensive ontology tree that had been developed and validated by several neuroradiologists in previous work (Buchlak et al., 2023). However, this initial ontology included compound terms (e.g., combinations of descriptors with observations or anatomy with observations) that were not directly suitable for our pipeline. We iteratively refined the ontology by examining head CT radiology reports from our dataset, simplifying it by removing these compound terms and adding additional entities as necessary. The resulting ontology accurately reflects the elemental findings present in head CT reports, including pathological observations, devices, and surgical changes. The resulting ontology tree can be seen in the Appendix.

Descriptor ontology: The descriptor ontology was developed by first establishing an initial schema based on an extensive literature review. We then analyzed radiology reports to identify and extract descriptor terms. Through an iterative refinement process aimed at reducing overlap and ensuring that distinct types of descriptors were captured appropriately, we arrived at a comprehensive and extendable descriptor ontology. The final version was reviewed and approved by our clinical team, and is presented on the Appendix.

Anatomy ontology: For anatomical entities, we use the Foundational Model of Anatomy (FMA)(Rosse and Mejino, 2003), a well-established ontology of anatomical knowledge. Our pipeline is

limited to entities accessible by starting with key anatomical structures relevant to head CTs (for example, the head and related structures such as the epidural space), with a maximum depth of 5 to avoid overly granular terms that are unlikely to appear in radiology reports.

3.3. Information Extraction Framework

We present a new information extraction framework that extends RadGraph (Jain et al., 2021) by introducing a distinct "descriptor" entity type. For clarity, we define key terms used in our framework as follows:

- **Observations**: Terms associated with visual radiological features, identifiable pathophysiologic processes, or diagnostic disease classifications (e.g., hypodensity, hemorrhage, hydrocephalus). Observations can be:
  - **Present (OBS-P)**: Indicates that an abnormal finding is reported as present in the radiology report (e.g., "there is a hemorrhage on the right").
  - **Absent (OBS-A)**: Indicates that an abnormal finding is explicitly reported as absent (e.g., "there is no hemorrhage").
- **Anatomy**: Refers to anatomical structures mentioned in the report, such as "frontal lobe" or "paranasal sinus."
- **Descriptors**: Modifier terms that describe either observations or anatomical entities (e.g., "large" for a hemorrhage; "dilated" for the lateral ventricles). For further details on the descriptor ontology, please see Appendix A.

Our framework retains RadGraph's anatomy and observation entities while adding this new category, enhancing the precision of information extraction from radiology reports. Our framework utilizes four relations: suggestive of, associated with, located at, and modify. To develop models capable of extracting information from head CT reports according to our schema, we use GPT-4 (1106-preview) (Achiam et al., 2023) to generate labeled entities and relations for a subset of our data. The prompts used for annotation are provided in the Appendix. Based on the annotated data, we train a Named Entity Recognition (NER) model using the Princeton University Relation Extraction (PURE) architecture (Zhong and Chen, 2021). This architecture employs a pipeline approach, decomposing the tasks of entity recognition and relation extraction into separate subtasks. For the NER model training, we follow RadGraph, use a learning rate of 1e-5 (tuning range 1e-4 to 1e-6) with a batch size of 16 for BERT and entity extraction, and use a learning rate of 2e-5 (tuning range 2e-4 to 2e-6) with a batch size of 16 for BERT for relation extraction. Our model achieved a precision of 0.88, recall of 0.89, and an F1 score of 0.88 for entity extraction, and a micro F1 score of 0.77 for relation extraction. We apply the trained model as the first step of HeadCT-ONE to extract all entities and relations from the reports.

3.4. Ontology Normalization

Following information extraction, we map all extracted entities to our predefined ontologies to ensure consistency and enable standardized analysis. For observation and anatomy entities, we leverage BioLORD-2023 (Remy et al., 2024), a state-of-the-art multilingual model, to generate high-fidelity embeddings of all ontology terms. For each extracted entity, we compute its embedding and compare it against the embeddings of all terms in the corresponding ontology. We then select the ontology term that has the highest similarity score to the entity's embedding, ensuring that the match is restricted to terms within the same entity type. We also apply some minor rule-based postprocessing pipelines to refine the extracted entities. For example, "frontoparietal" would be separated into "frontal" and "parietal" to align with our ontology structure. For descriptors, given their limited categorical nature, we employ GPT-4o (version gpt4o05132024) to classify extracted entities into our predefined descriptor ontology categories using a prompt provided in the Appendix. This normalization process facilitates consistent and controllable analyses across diverse reports.

3.5. HeadCT-ONE Metric

HeadCT-ONE metric assesses the model's ability to correctly identify entities (including the new descriptor type) and their relations in radiological reports. The HeadCT-ONE metric is calculated as the average of two weighted F1 scores: one for entity recognition and one for relation extraction. Weights are assigned to individual items (entities or relations) based on criteria such as entity type, relation type, and clinical relevance. This weighting mechanism allows the metric to be tailored to specific evaluation needs.

Metric	Weight	Normal	Normal-Abnormal
GREEN	✗	0.691	0.174
RaTEScore	✗	0.713	0.421
RadGraph F1	✗	0.328	0.160
HeadCT-ONE	✗	0.429	0.169
RadGraph F1	✓	0.224	0.149
HeadCT-ONE	✓	**0.903**	**0.903**

Table 1: Comparison across 20 sites for (1) average metric scores when comparing pairs of normal reports from different sites and (2) the difference between these scores and those obtained when comparing normal and abnormal reports from different sites. The "Weight" column indicates whether a weighting scheme was used. Each site contributed one normal report and one abnormal report in this analysis.

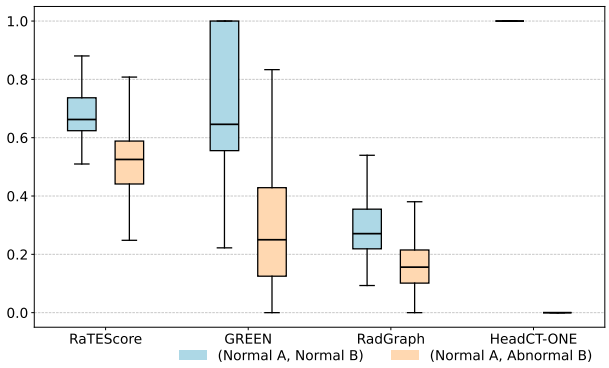

Figure 2: Scores of different report generation metrics when comparing two normal reports (light blue) or one normal and one abnormal report (light orange).

Define e_{gt} and e_{pred} as the sets of ground truth and predicted entities, and r_{gt} and r_{pred} be the sets of ground truth and predicted relations, respectively. The metric is calculated as the average of two weighted F1 scores:

$$\text{HeadCT-ONE} = \frac{F1(e_{gt}, e_{pred}) + F1(r_{gt}, r_{pred})}{2},$$

where $F1$ is a weighted F1 score calculated based on precision (P) and recall (R).

We conducted inference on a Tesla V100 GPU (16 GB memory), achieving an average inference time of 5 seconds per reference-candidate pair for HeadCT-ONE calculation; however, only about 2 GB of memory is required, and the process can also be run on a CPU, albeit with a longer runtime.

4. Results

4.1. Analysis on Original Reports

Ontology normalization enables HeadCT-ONE to consistently identify similar concepts across diverse radiology reports. To evaluate this, we conducted a comparative analysis using reports from 20 different sites. For each site, one normal report and one abnormal report were selected. Our experiment tested how well various metrics could recognize similarities between normal reports and distinguish between normal and abnormal reports, regardless of differences in writing styles.

Table 1 presents the average scores for these comparisons. For each site, we computed a "Normal" score by comparing its normal report with normal reports from other sites, and a "Normal-Abnormal" score by calculating the difference between the previously mentioned normal vs normal comparison, and comparisons between normal reports from each site to abnormal reports from other sites. This "Normal-Abnormal" score effectively measures each metric's ability to differentiate between normal and abnormal reports while avoiding biases from consistently high or low scores. Additionally, Figure 2 presents a boxplot visualization of results across 20 sites, providing insight into score distribution.

Our results demonstrate that HeadCT-ONE consistently outperformed RadGraph F1, achieving higher scores across all metrics. This superiority suggests that normalizing the data using ontologies significantly improves the overall analysis performance.

Weighting on OBS-P enables HeadCT-ONE to easily distinguish between normal and abnormal reports. As illustrated in Table 1 and Figure 2, after adding weights to all entities and relations related to observation-present, HeadCT-ONE demonstrated an improved ability to distinguish between normal and abnormal reports. This is evidenced by HeadCT-ONE achieving the highest score in the (Normal A, Normal B) - (Normal A, Abnormal B) comparison. This indicates that the weighted approach in HeadCT-ONE is more sensitive to the differences between normal and abnormal reports.

Metric	Weight				F1 Score								
	OBS-P	OBS-A	ANAT	DESC	Reph	Any	Obs	Ana	Des	Reph-Any.	Reph-Obs	Reph-Ana	Reph-Des
GREEN	/	/	/	/	0.820	0.639	0.707	0.703	0.679	0.182	0.114	0.118	0.142
RaTEScore	/	/	/	/	0.784	0.733	0.730	0.759	0.744	0.051	0.054	0.025	0.040
RadGraph F1	1	1	1	/	0.426	0.341	0.349	0.343	0.354	0.085	0.077	0.083	0.072
	1	0	0	/	0.393	0.334	0.328	0.380	0.307	0.059	0.065	0.013	0.086
	1	1	0	/	0.360	0.329	0.324	0.354	0.310	0.032	0.036	0.006	0.050
	0	0	1	/	0.516	0.448	0.467	0.419	0.497	0.068	0.050	0.097	0.019
HeadCT-ONE	1	1	1	1	0.310	0.274	0.283	0.279	0.274	0.037	0.027	0.031	0.036
w/o Ontology	1	0	0	0	0.440	0.241	0.237	0.397	0.313	0.199	0.203	0.044	0.128
	1	1	0	0	0.243	0.217	0.215	0.230	0.221	0.027	0.029	0.013	0.023
	0	0	1	0	0.346	0.287	0.319	0.255	0.327	0.060	0.028	0.091	0.019
	0	0	0	1	0.342	0.311	0.314	0.337	0.281	0.032	0.029	0.006	0.062
HeadCT-ONE	1	1	1	1	0.631	0.558	0.571	0.581	0.571	0.073	0.060	0.051	0.060
	1	0	0	0	0.719	0.477	0.468	0.657	0.570	**0.242**	**0.251**	0.062	**0.149**
	1	1	0	0	0.671	0.597	0.597	0.645	0.625	0.074	0.074	0.026	0.046
	0	0	1	0	0.544	0.446	0.495	0.407	0.512	0.099	0.050	**0.138**	0.032
	0	0	0	1	0.646	0.590	0.587	0.635	0.554	0.056	0.060	0.011	0.092

Table 2: Comparison of F1 scores for various metrics and weighting schemes applied to 400 radiology reports, including original, rephrased, and error-inserted versions. Weighting schemes (OBS-P: Observation-present, OBS-A: Observation-absent, ANAT: Anatomy, DESC: Descriptor) use 1 or 0 to indicate the weighing number of related entities and relations. Columns show scores for rephrased reports (Reph) and reports with errors of any type (Any), with observation errors (Obs), anatomy errors (Ana) and descriptor errors (Des). The last four columns (Reph-Any, Reph-Obs, Reph-Ana, Reph-Des) represent score differences between (Original, Rephrased) and (Original, Error-Inserted Version).

The effectiveness of this weighting strategy is further demonstrated by HeadCT-ONE's performance on normal pairs, with a perfect score of 1 for normal pairs in 19 out of 20 sites. The one site where HeadCT-ONE did not achieve a perfect score was due to an error assignment of "aeration" as "observation_present" in the sentence "The bone windows demonstrate normal aeration of the paranasal sinuses and mastoid air cells", leading to this discrepancy. This example suggests that there is room for further refinement in its entity recognition accuracy, particularly for terms that may have nuanced meanings depending on their context.

4.2. Analysis on Modified Reports

The weighting mechanism for entity types allows HeadCT-ONE to focus on specific aspects of AI-generated reports. Before presenting Table 2, we note that for this analysis we use a binary weighting scheme: a weight of "1" indicates that an entity or relation is fully included in the weighted F1 score calculation, while a weight of "0" excludes it. These weights are fixed for our analysis (i.e., assigned as 1 or 0), although they can be adjusted based on the task at hand. HeadCT-ONE's weighting mechanism for entity types provides a powerful tool for customizing the analysis of AI-generated radiology reports. As demonstrated in Table 2, this allows us to prioritize the detection of specific error types by adjusting the weights assigned to different entity categories. For example, when there is a particular concern about observation-related errors, increasing the weight of the observation-present (OBS-P) category enhances HeadCT-ONE's sensitivity to these types of discrepancies. This is evident in the results where the weighting scheme (1, 0, 0, 0) for (OBS-P, OBS-A, ANAT, DESC) yields the highest F1 score differences for observation errors (Reph-Obs: 0.251) and general errors (Reph-Any: 0.242). Similarly, if the focus is on detecting anatomy-related errors, assigning a higher weight to the anatomy (ANAT) category improves performance in this area. The weighting scheme (0, 0, 1, 0) results in the highest F1 score difference for anatomy errors (Reph-Ana: 0.149), demonstrating HeadCT-ONE's ability to adapt to this specific focus.

Report	Extracted Entities Related to Observation-Present
Report1: The patient is status post left frontotemporal craniotomy and aneurysm clipping. Small areas of encephalomalacia within the left temporal lobe, unchanged. No hemorrhage, hydrocephalus, mass lesion, acute infarct, or extra-axial fluid collection. Visualized calvarium unremarkable.	left frontotemporal -> Left side of frontal bone, Left temporal part of head craniotomy -> craniotomy aneurysm clipping -> vascular clips small -> small encephalomalacia -> encephalomalacia left temporal lobe -> Left temporal lobe unchanged -> stable
Report2: The patient has undergone a left frontotemporal craniotomy and aneurysm clipping. There are stable, small regions of encephalomalacia within the left temporal lobe. There is a new small region of acute hemorrhage in the left frontal lobe. No evidence of hydrocephalus, neoplastic process, acute cerebral infarction, or extra-axial fluid accumulation is observed. The visualized portions of the calvarium appear unremarkable.	left frontotemporal -> Left side of frontal bone, Left temporal part of head craniotomy -> craniotomy aneurysm clipping -> vascular clips stable -> stable small -> small encephalomalacia -> encephalomalacia left temporal lobe -> Left temporal lobe new -> new acute -> acute hemorrhage -> hemorrhage left frontal lobe -> Left frontal lobe
Report3: Postop changes from right frontoparietal craniotomy again seen. Small right-sided subdural hematoma and extra-axial gas remain stable measuring approximately 13 mm in thickness. Mild right to left midline shift is also stable. No new or increased areas of intracranial hemorrhage are seen. No evidence of brain edema or other signs of acute infarction. Ventricles stable in size.	right frontoparietal -> Right side of frontal bone, Right parietal part of head craniotomy -> craniotomy small -> small right-sided -> right-sided of head subdural -> subdural space hematoma -> hemorrhage extra-axial -> anterior to posterior gas -> gas stable -> stable 13 mm -> numeric mild -> mild midline shift -> midline shift

[ANATOMY] [DESCRIPTOR] [OBSERVATION_PRESENT] [OBSERVATION_ABSENT]

No Weights			Weights on OBS-P			Weights on Hemorrhage (OBS-P)		
(R1, R2)	(R1, R3)	(R2, R3)	(R1, R2)	(R1, R3)	(R2, R3)	(R1, R2)	(R1, R3)	(R2, R3)
0.63	0.27	0.28	0.82	0.16	0.33	0.00	0.00	0.61

Figure 3: Case study. F1 scores between report pairs (R1,R2), (R1,R3), and (R2,R3) under different weighting schemes. No weights shows baseline similarities. Weights on OBS-P (observation-present) increases similarity for reports sharing many positive observations. Weights on 'Hemorrhage' (OBS-P) highlights specific differences in hemorrhage mentions, drastically changing similarities between otherwise similar reports.

The weighting of specific entities enables precise identification of targeted error types in radiology reports, enhancing its controllability. The application of entity-specific weights in HeadCT-ONE demonstrates the ability to focus on particular aspects of radiology reports, thereby improving error detection in targeted areas. As shown in Figure 3, when applying weights to any observation-present entities, two reports sharing many positive observations except for hemorrhage (R1, R2) score high (0.82). However, this score drops to 0 when weighting only on specific 'hemorrhage' observation-present entities, highlighting their key difference. Conversely, a different pair of reports with more distinct observations but both mentioning different types of hemorrhage (R2, R3) scores lower (0.33) when weighting on any observation-present entities. This pair's score increases to 0.61 when weighting specifically on hemorrhage, though it doesn't reach 1.0 due to the different types of hemorrhage mentioned.

4.3. Alignment with Radiologists

We investigate the correlation between automated metrics and radiologist assessments of radiology reports to validate the effectiveness of the HeadCT-ONE metric. A senior radiology resident classified modified reports as with and without clinically significant errors (i.e., errors that could impact patient care) compared to candidate reports. We then calculated the scores for various metrics on modified reports vs. candidate reports, The results are presented in Table 3. We explore HeadCT-ONE with weighting only on observation-present entities ("OBS-P") and applying higher weight to the top 5 or top 10 most frequently negated observation-present entities ("Top 5" and "Top 10" respectively). The results demonstrate that HeadCT-ONE exhibits promising performance

Metric	Weight	w/o Sig.	w Sig.	Diff.
GREEN	✗	0.743	0.688	0.055**
RaTEScore	✗	0.750	0.744	0.006*
RadGraph F1	✗	0.366	0.371	-0.005*
HeadCT-ONE	OBS-P	0.716	0.543	0.173
HeadCT-ONE	Top 5	0.732	0.497	0.235
HeadCT-ONE	Top 10	0.706	0.409	**0.297**

Table 3: Scores for metrics on reports with and without clinically significant errors (Sig.). Clinically significant errors are errors that could impact patient care. 'Diff.' shows the difference between 'w/o Sig.' and 'w Sig.'. 'Weight' indicates whether the metric uses weighting. * $p < 0.05$, ** $p < 0.01$ compared to HeadCT-ONE (Top 10).

in distinguishing between reports with and without significant errors. Notably, the "Top 10" weighting strategy for HeadCT-ONE shows the largest difference (0.297) between scores for reports without significant errors (0.706) and those with significant errors (0.409). This analysis validates HeadCT-ONE's capacity to capture clinically relevant aspects of report quality and highlights the potential for adjusting the metric to better align with radiologist judgments.

5. Discussion

Our study introduces HeadCT-ONE, an ontology-enhanced information extraction-derived metric for head CT radiology report generation. Our results showcase that HeadCT-ONE offers significant improvements over existing metrics in terms of robustness, granularity, and controllability.

Entity normalization enhances information extraction-derived metrics. Medical ontologies have a widespread use in healthcare, but their integration into radiology report generation and evaluation is challenging. Widely used ontologies such as SNOMED-CT (Gaudet-Blavignac et al., 2021), International Classification of Diseases (ICD) (Who, 2005) and Unified Medical Language System (UMLS) (Bodenreider, 2004) do not generalize well to the language utilized in radiology reports. While RadLex (Chepelev et al., 2023) aims to address radiology-specific terminology, its scope may not fully encompass all subspecialty terms (Datta et al., 2020), and its granularity can be inconsistent, sometimes lacking detail for certain findings while being overly specific for others, potentially affecting standardized usage across different radiologists and institutions. In our study, we modified an ontology tree specifically designed for head CT findings, and developed an ontology for radiology descriptors based on head CT but easily extendable to other CT study types. In addition, we use the FMA for anatomical ontology as we found it to match radiological terms better than other ontologies. In our experiments, we show that the normalization step improves the capability of HeadCT-ONE to identify similar reports regardless of variations of radiological language.

Entity weighting allows for controllable evaluation and better alignment with experts. Radiology reports encompass a complex hierarchy of information, where the clinical significance of elements varies across different study types and analytical objectives. For instance, when focusing on diagnostic accuracy, the presence or absence of critical observations may be paramount, while the comprehensive use of descriptors would be relevant when evaluating report completeness. This variability extends to specific clinical scenarios: in a head trauma context, the evaluation might prioritize acute findings such as intracranial hemorrhage, midline shift, or skull fractures. Conversely, in the longitudinal follow-up of hydrocephalus, the focus may shift to changes in ventricular size and shunt positioning. This multifaceted variability in the importance of report elements poses a significant challenge for traditional evaluation metrics, which often treat all terms with equal weight. Our pipeline introduces a weighting mechanism that allows for focused evaluation on broad or specific entity types. This controllability enables more nuanced and clinically relevant assessments, as evidenced by the improved correlation with radiologist evaluations. Moreover, in recognizing the current limitations of generative AI in radiology, our approach allows for strategic leniency in certain areas. For instance, one could modulate the impact of discrepancies in quantitative descriptors like precise measurements, which AI systems may struggle to consistently reproduce.

Limitations. Our study is not without limitations. The normalization process, while effective, may occasionally misclassify terms. Further, our analyses were done on synthetic reports, which may not completely represent outputs from real head CT report generation models. Additionally, the creation of comprehensive ontologies for different radiological domains re-

mains a challenging and time-consuming task requiring domain knowledge. Looking ahead, promising directions for future research include combining our information extraction and normalization approach with the increasing capabilities of LLMs. LLMs have shown to perform well at information extraction (Liu et al., 2023). For normalization, LLMs could provide more sophisticated and accurate term classification, enhancing the overall reliability of HeadCT-ONE. Additionally, leveraging LLMs for data-driven ontology creation could significantly improve scalability, reducing the need for extensive domain knowledge and expert involvement. Furthermore, as accurate head CT report generation models become available, validating HeadCT-ONE on their outputs will be crucial to ensure its effectiveness in evaluating real-world generated reports. Additionally, we acknowledge that the current assessment of clinically significant errors was conducted by a single radiology resident, which may introduce subjectivity. In our planned future work, we plan to involve a panel of board-certified radiologists to quantify inter-rater reliability and further validate our metric's alignment with clinical preferences.

In conclusion, HeadCT-ONE enhances AI-generated head CT report evaluation, offering improved robustness, granularity, and controllability, being easily extendable to other imaging modalities.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Oishi Banerjee, Agustina Saenz, Kay Wu, Warren Clements, Adil Zia, Dominic Buensalido, Helen Kavnoudias, Alain S. Abi-Ghanem, Nour El Ghawi, Cibele Luna, Patricia Castillo, Khaled Al-Surimi, Rayyan A. Daghistani, Yuh-Min Chen, Heng sheng Chao, Lars Heiliger, Moon Kim, Johannes Haubold, Frederic Jonske, and Pranav Rajpurkar. Rexamine-global: A framework for uncovering inconsistencies in radiology report generation metrics, 2024. URL https://arxiv.org/abs/2408.16208.

Olivier Bodenreider. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res.*, 32(Database issue):D267–70, January 2004.

Quinlan D Buchlak, Cyril H M Tang, Jarrel C Y Seah, Andrew Johnson, Xavier Holt, Georgina M Bottrell, Jeffrey B Wardman, Gihan Samarasinghe, Leonardo Dos Santos Pinheiro, Hongze Xia, Hassan K Ahmad, Hung Pham, Jason I Chiang, Nalan Ektas, Michael R Milne, Christopher H Y Chiu, Ben Hachey, Melissa K Ryan, Benjamin P Johnston, Nazanin Esmaili, Christine Bennett, Tony Goldschlager, Jonathan Hall, Duc Tan Vo, Lauren Oakden-Rayner, Jean-Christophe Leveque, Farrokh Farrokhi, Richard G Abramson, Catherine M Jones, Simon Edelstein, and Peter Brotchie. Effects of a comprehensive brain computed tomography deep learning model on radiologist detection accuracy. *Eur. Radiol.*, August 2023.

Juan Manuel Zambrano Chaves, Shih-Cheng Huang, Yanbo Xu, Hanwen Xu, Naoto Usuyama, Sheng Zhang, Fei Wang, Yujia Xie, Mahmoud Khademi, Ziyi Yang, Hany Awadalla, Julia Gong, Houdong Hu, Jianwei Yang, Chunyuan Li, Jianfeng Gao, Yu Gu, Cliff Wong, Mu Wei, Tristan Naumann, Muhao Chen, Matthew P. Lungren, Akshay Chaudhari, Serena Yeung-Levy, Curtis P. Langlotz, Sheng Wang, and Hoifung Poon. Towards a clinically accessible radiology foundation model: open-access and lightweight, with automated evaluation, 2024. URL https://arxiv.org/abs/2403.08002.

Leonid L Chepelev, David Kwan, Charles E Kahn, Ross W Filice, and Kenneth C Wang. Ontologies in the new computational age of radiology: RadLex for semantics and interoperability in imaging workflows. *Radiographics*, 43(3):e220098, March 2023.

Surabhi Datta, Jordan Godfrey-Stovall, and Kirk Roberts. RadLex normalization in radiology reports. *AMIA Annu. Symp. Proc.*, 2020:338–347, 2020.

Mark Endo, Rayan Krishnan, Viswesh Krishna, Andrew Y. Ng, and Pranav Rajpurkar. Retrieval-based chest x-ray report generation using a pre-trained contrastive language-image model. In Subhrajit Roy, Stephen Pfohl, Emma Rocheteau, Girmaw Abebe Tadesse, Luis Oala, Fabian Falck, Yuyin Zhou, Liyue Shen, Ghada Zamzmi, Purity Mugambi, Ayah Zirikly, Matthew B. A. McDermott, and Emily Alsentzer, editors, *Proceedings*

of Machine Learning for Health, volume 158 of *Proceedings of Machine Learning Research*, pages 209–219. PMLR, 04 Dec 2021. URL https://proceedings.mlr.press/v158/endo21a.html.

Christophe Gaudet-Blavignac, Vasiliki Foufi, Mina Bjelogrlic, and Christian Lovis. Use of the systematized nomenclature of medicine clinical terms (SNOMED CT) for processing free text in health care: Systematic scoping review. *J. Med. Internet Res.*, 23(1):e24594, January 2021.

Stephen Gilbert, Jakob Nikolas Kather, and Aidan Hogan. Augmented non-hallucinating large language models as medical information curators. *NPJ Digit. Med.*, 7(1):100, April 2024.

Alyssa Huang, Oishi Banerjee, Kay Wu, Eduardo Pontes Reis, and Pranav Rajpurkar. Fineradscore: A radiology report line-by-line evaluation technique generating corrections with severity scores, 2024. URL https://arxiv.org/abs/2405.20613.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison, 2019. URL https://arxiv.org/abs/1901.07031.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P. Lungren, Andrew Y. Ng, Curtis P. Langlotz, and Pranav Rajpurkar. Radgraph: Extracting clinical entities and relations from radiology reports, 2021. URL https://arxiv.org/abs/2106.14463.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

Qianchu Liu, Stephanie Hyland, Shruthi Bannur, Kenza Bouzid, Daniel Castro, Maria Wetscherek, Robert Tinn, Harshita Sharma, Fernando Pérez-García, Anton Schwaighofer, Pranav Rajpurkar, Sameer Khanna, Hoifung Poon, Naoto Usuyama, Anja Thieme, Aditya Nori, Matthew Lungren, Ozan Oktay, and Javier Alvarez-Valle. Exploring the boundaries of GPT-4 in radiology. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14414–14445, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.891. URL https://aclanthology.org/2023.emnlp-main.891.

Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, April 2023.

Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson, Michael Moseley, Curtis Langlotz, Akshay S Chaudhari, and Jean-Benoit Delbrouck. Green: Generative radiology report evaluation and error notation, 2024. URL https://arxiv.org/abs/2405.03595.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

Pranav Rajpurkar and Matthew P Lungren. The current and future state of AI interpretation of medical images. *N. Engl. J. Med.*, 388(21):1981–1990, May 2023.

François Remy, Kris Demuynck, and Thomas Demeester. Biolord-2023: semantic textual representations fusing large language models and clinical knowledge graph insights. *Journal of the American Medical Informatics Association*, page ocae029, 2024.

Cornelius Rosse and José L V Mejino, Jr. A reference ontology for biomedical informatics: the foundational model of anatomy. *J. Biomed. Inform.*, 36 (6):478–500, December 2003.

Phillip Sloan, Philip Clatworthy, Edwin Simpson, and Majid Mirmehdi. Automated radiology report generation: A review of recent advances. *IEEE Reviews in Biomedical Engineering*, page 1–20, 2024. ISSN 1941-1189. doi: 10.1109/rbme.2024.3408456. URL http://dx.doi.org/10.1109/RBME.2024.3408456.

Who. *The international statistical classification of diseases and health related problems ICD-10: Tabular list v. 1: Tenth revision*. World Health Organization, Genève, Switzerland, 2 edition, April 2005.

Yiqing Xie, Sheng Zhang, Hao Cheng, Pengfei Liu, Zelalem Gero, Cliff Wong, Tristan Naumann, Hoifung Poon, and Carolyn Rose. Doclens: Multi-aspect fine-grained evaluation for medical text generation, 2024. URL https://arxiv.org/abs/2311.09581.

Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y Ng, Curtis P Langlotz, Vasantha Kumar Venugopal, and Pranav Rajpurkar. Evaluating progress in automatic chest x-ray radiology report generation. *Patterns (N Y)*, 4(9):100802, September 2023.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert, 2020. URL https://arxiv.org/abs/1904.09675.

Weike Zhao, Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Ratescore: A metric for radiology report generation, 2024. URL https://arxiv.org/abs/2406.16845.

Zexuan Zhong and Danqi Chen. A frustratingly easy approach for entity and relation extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 50–61, 2021.

Appendix A. Defined ontologies

Table A1: Descriptors ontology categorization, showing examples for each category. The first level represents the main category, followed by a more specific second and third level where applicable.

First level	**Second level**	**Third level**	**Examples**
quantity	numeric		5 lesions,3 fractures
	qualitative	single	isolated,solitary,single
		multiple	several,multiple,numerous,a few
size	numeric		5 mm,10 cm
	qualitative	very_small	tiny,microscopic
		small	small
		medium	moderate,average,medium size,normal size
		large	large,big,enlarged
		very_large	enormous,huge,very large,gigantic,very enlarged
shape	regular	spherical	round,oval,ovoid
		saccular	saccular
		curvilinear	curvilinear
		crescentic	crescentic
		biconvex	biconvex
		laminar	laminar,sheet-like,layer
		tubular	tubular,cylindrical
		fusiform	fusiform
	irregular	lobulated	lobulated
		spiculated	spiculated
		amorphous	amorphous
homogeneity	homogeneous		homogeneous
	heterogeneous		heterogeneous
density	hypodense		hypodense,hypoattenuation,hypodensity
	isodense		isodense
	hyperdense		hyperdense,hyperattenuation,hyperdensity
	mixed		mixed density
margin	well_defined		circumscribed,well defined,well circumscribed,well delimited
	poorly_defined		ill-defined,poorly circumscribed
severity	minimal		minimal
	mild		mild
	moderate		moderate
	severe		severe
temporality	acute		acute,new
	subacute		subacute
	chronic		chronic,old,remote
	acute_on_chronic		acute on chronic

Continued on next page

Table A1 – continued from previous page

First level	Second level	Third level	Examples
	age_indeterminate		age-indeterminate,unknown age
distribution	localized		focal,localized
	diffuse		diffuse
	confluent		confluent
	scattered		scattered
	petechial		petechial
	multifocal		multifocal
enhancement	present	homogeneous	homogeneous enhancement
		heterogeneous	heterogeneous enhancement
		peripheral	peripheral enhancement
		central	central enhancement
		rim	rim-like enhancement
		patchy	patchy enhancement
	absent		no enhancement
certainty	definitely_present		there is, there are,with
	probably_present		probably,likely
	possibly_present		possibly
	uncertain		cannot rule out
	definitely_absent		no evidence of,there is no,without
composition	gas		gas,gaseous,air
	fluid	simple_fluid	fluid-like,simple fluid
		csf	csf
		serous	serous
		hemorrhagic	hemorrhagic
		mucinous	mucinous,colloid
	solid	soft_tissue	soft-tissue density
		fatty	fatty
		fibrous	fibrous
		calcified	calcific density,calcified
		sclerotic	sclerotic
	mixed		mixed,semisolid,fluid and solid components,solid with hemorrhagic components
complexity	simple		simple
	complex		complex
change	resolution		resolution,resolved,cleared,disappeared
	improvement		improved,improving
	increase		increased,increase in size,larger,increasing
	decrease		decreased,decrease in size,decreasing,smaller
	worsening		worsened,worsening
	appearance		new,appeared,is now present
	mixed_change		one metastasis increased in size and the other one resolved
	stable		stable,unchanged,similar,similar in appearance

Continued on next page

Table A1 – continued from previous page

First level	Second level	Third level	Examples
normalcy	normal		normal
	abnormal		abnormal
caliber	dilated		dilated ventricles,dilated vessels
	normal		average lumen,normal cavity,normal calibre
	reduced		narrowed artery,collapsed veins,collapsed ventricles
malignancy_status	definitely_benign		benign
	probably_benign		probably benign lesion
	indeterminate		indeterminate hypodensity
	probably_malignant		probably malignant,suspicious
	definitely_malignant		cancerous,malignant
patency	patent		patent,clear
	mostly_patent		mostly patent
	obstructed		obstruction,obstructed
	occluded		occlusion
occupancy	empty		empty,clear,well-aereated
	partially_filled		partially filled with..
	fully_filled		full,fully filled
	engorged		engorged
integrity	intact		intact,unruptured
	partially_compromised		partially ruptured, partially disrupted
	compromised		ruptured,disrupted,disruption
direction	left_to_right		left-to-right
	right_to_left		right-to-left
	anterior_to_posterior		anterior-to-posterior
	posterior_to_anterior		posterior-to-anterior
	upwards		upwards
	downwards		downwards
component_involved	mucosal		mucosal
	muscular		muscular
	osseous		osseous,bony
position	normal_position		normal position
	abnormal_position		displaced,abnormal position

Table A2: Findings ontology list, with synonyms when relevant.

Finding	**Synonyms**
infarct	ischemic stroke,infarction
hemorrhage	hematoma,bleed,blood
agenesis	
lesion	mass,tumor,tumour
thickening	
aneurysm	
coils	
subluxation	
dissociation	
effacement	
calcification	
thrombosis	clot,thrombus
beam_hardening_artefact	
atrophy	involution,atrophic changes
cavum_septum_pellucidum	
chiari_1	
chiari_2	
cochlear_implant	
cyst	
colpocephaly	
hydrocephalus	
hypodensity	
necrosis	
craniotomy	
collection	
dbs_electrodes	
thinning	
demyelination	
diffuse_axonal_injury	
venous_gas	
arachnoidocele	
encephalitis	
encephalomalacia	
entrapment	
external_ventricular_drainage	ventriculostomy catheter,ventriculostomy
exophthalmos	
empyema	
herniation	
fracture	
erosion	
fibrous_dysplasia	
foreign_body	
fungal_sinusitis	
shape_abnormality	
heterotopia	
hyperdense_artery	

Continued on next page

Table A2 – continued from previous page

Finding	Synonyms
hyperostosis	
hypopneumatisation	
hypoxic_ischaemic_encephalopathy	
intracranial_pressure_monitor	icp
insular_ribbon_sign	
silicone	
debris	
opacity	
post_surgical_change	
meningioma	
metallic_artefact	
midline_shift	
movement_artefact	
mucocoele	
non_hemorrhagic_contusion	
optic_neuritis	
abscess	
fat_stranding	
prosthesis	
osteoma	
otosclerosis	
papilloedema	
perivascular_spaces	
pseudo_sah	
resection_cavity	
schizencephaly	
haemangioma	
small_vessel_disease	white matter change,white matter changes,ischemic change,ischemic changes,microvascular changes,microvascular disease,microvascular change
stapes_implants	
ectopic_air	emphysema,pneumocephalus
arthritis	
dislocation	
edema	
transphenoidal_surgery	
vascular_clips	aneurysm clips
vascular_stents	stent,stents
venous_infarct	
venous_thrombosis	cvt,venous sinus thrombosis,cerebral venous thrombosis
ventriculoperitoneal_shunt	vp shunt
mass_effect	
loss_of_gray_white_matter_differentiation	
abnormality	pathology,finding,process,abnormalities

Appendix B. Prompt for rephrasing and error introduction

```
You are an expert radiologist tasked with rephrasing a given radiology report. Your objective is to rewrite the report using alternative medical terminology and sentence structures while maintaining the accuracy of the medical content. The rewritten report should sound distinctly different from the original but convey the same clinical information in a professional radiologist's style.

Please adhere to the following guidelines:
1. Utilize synonymous medical terms and phrases where appropriate.
2. Restructure sentences to present information in a different order or format.
3. Ensure that all key medical findings and diagnoses are preserved in the rewritten version.
4. Adapt the level of detail to match the original report.

Here's an example to illustrate the task:

Original Report:
FINDINGS: The lungs are clear without focal consolidation. No pleural effusion or pneumothorax is seen. The heart size is normal. The mediastinum is unremarkable.

Rewritten Report:
FINDINGS: Pulmonary fields demonstrate no evidence of focal opacities or airspace disease. Pleural spaces are free of fluid collections or air. Cardiac silhouette is within normal limits. Mediastinal structures appear unremarkable.

Please provide the rewritten report, maintaining medical accuracy while changing the wording and phrasing.

Original Radiology Report:
{report}
Rewritten Report:
```

```
You are an expert radiologist tasked with modifying a given radiology report by introducing 1-3 errors while maintaining overall medical coherence. Your objective is to rewrite the report with subtle yet significant changes that could impact the interpretation of the findings.

Definitions:
- observation: A specific finding or abnormality noted in a radiology report. An observation typically describes a particular condition, anomaly, or feature identified during the imaging study.
- anatomical location: The specific area or structure in the body an observation is found, or an anatomical structure being described.
- descriptor: A characteristic or attribute used to describe an observation, such as size, shape, density, or severity.

Please adhere to the following guidelines:
1. Introduce 1-3 errors from any of the following categories:
   a) Observation errors:
      - Missing an abnormal observation: Omit an abnormal finding that was present in the original report.
      - Added an abnormal observation: Include a new abnormal finding that was not present in the original report.
      - Negated an abnormal observation: Change an abnormal finding to a negative or normal statement.
   b) Anatomical errors:
      - Incorrect anatomical location: Change the location of an observation to a different but plausible anatomical site, lobe, or region.
      - Incorrect anatomical side: Switch the side (left/right) of an observation.
   c) Descriptor errors:
      - Change one or more aspects of an observation, such as quantity, size, shape, homogeneity, density, margin, severity, temporality, distribution, enhancement, temporal
```

```
     change, certainty, composition, complexity, normalcy, caliber, malignancy status,
    patency, occupancy, integrity, direction, component involved, or position.
2. Ensure the errors are plausible and maintain medical coherence within the context of the
    report. You can modify the rest of the report to be consistent with the introduced
    errors.
3. Preserve the overall structure and style of the original report.

Please provide the modified report with introduced errors while maintaining overall medical
    coherence. Additionally, include information about which errors were introduced.

Your response should be in JSON format with two main keys: "modified_report" for the
    rewritten report, and "errors" for describing the introduced errors.

Original Radiology Report:
{report}

Please provide the rewritten report with 1-3 errors (observation, anatomical, or descriptor)
     in JSON format. Use the following structure:

{
    "modified_report": "Your rewritten report here",
    "errors": [
        {
            "type": "error type: observation (missing/added/negated), anatomical (incorrect
    location/incorrect side), or descriptor (specify type)",
            "description": "brief description of the error"
        },
        ...
    ]
}
```

```
You are an expert radiologist tasked with modifying a given radiology report by introducing
    1-3 observation errors while maintaining overall medical coherence.
Your objective is to rewrite the report with subtle yet significant changes that could
    impact the interpretation of the findings.

Definitions:
- observation: A specific finding or abnormality noted in a radiology report. An observation
     typically describes a particular condition, anomaly, or feature identified during the
    imaging study.
- negation or normal statement: an statement commenting on the absence of a particular
    abnormal finding, or the normalcy of a structure.

Please adhere to the following guidelines:
1. Introduce 1-3 observation errors, which can be one of the following types:
   a) Missing an abnormal observation: Omit an abnormal finding that was present in the
    original report. DO NOT remove negative or normal statements.
   b) Added an abnormal observation: Include a new abnormal finding that was not present in
    the original report. DO NOT add negative or normal statements.
   c) Negated an abnormal observation: Change a abnormal finding to a negative or normal
    statement using the appropriate general negation or normal statements.
2. Avoid making changes to the anatomical location of observations (anatomical errors) and
    descriptors of these observations (e.g., severity).
3. Ensure the errors are plausible and maintain medical coherence within the context of the
    report. You can modify the rest of the report to be consistent with that error. The
    report should still make sense medically, even with the introduced errors.
4. Preserve the overall structure and style of the original report, even in the errors
    introduced.

Please provide the modified report with introduced observation errors while maintaining
    overall medical coherence. Additionally, include information about which errors were
    introduced.

```

```
Your response should be in JSON format with two main keys: "modified_report" for the rewritten report, and "errors" for describing the introduced errors.

Original Radiology Report:
{report}

Please provide the rewritten report with 1-3 observation errors in JSON format. Use the following structure:

{
    "modified_report": "Your rewritten report here",
    "errors": [
        {
            "type": "error type: missing/added/replaced/negated",
            "description": "brief description of the error"
        },
        ...
    ]
}
```

```
You are an expert radiologist tasked with modifying a given radiology report by introducing 1-3 anatomical errors while maintaining overall medical coherence.
Your objective is to rewrite the report with subtle yet significant changes that could impact the interpretation of the findings.

Definitions:
- anatomical error: An error in the description of the anatomical location of structures or findings in a radiology report.
- observation: A specific finding or abnormality noted in a radiology report. An observation typically describes a particular condition, anomaly, or feature identified during the imaging study.

Please adhere to the following guidelines:
1. Introduce 1-3 anatomical errors, which can be one of the following types:
   a) Incorrect anatomical location: Change the location of an observation to a different but plausible anatomical site, lobe or region for this particular observation, and study type.
   b) Incorrect anatomical side: Switch the side (left/right) of an observation.
2. Avoid making changes to the observations themselves (observation errors) and descriptors of these observations (e.g., severity).
3. Ensure the errors are plausible and maintain medical coherence within the context of the report. You can modify the rest of the report to be consistent with that error. The report should still make sense medically, even with the introduced errors.
4. Preserve the overall structure and style of the original report, even in the errors introduced.

Please provide the modified report with introduced anatomical errors while maintaining overall medical coherence. Additionally, include information about which errors were introduced.

Your response should be in JSON format with two main keys: "modified_report" for the rewritten report, and "errors" for describing the introduced errors.

Original Radiology Report:
{report}

Please provide the rewritten report with 1-3 anatomical errors in JSON format. Use the following structure:

{
    "modified_report": "Your rewritten report here",
    "errors": [
        {
            "type": "error type: incorrect location/incorrect side",
```

```
            "description": "brief description of the error"
        },
        ...
    ]
}
```

```
You are an expert radiologist tasked with modifying a given radiology report by introducing 1-3 descriptor errors while maintaining overall medical coherence.
Your objective is to rewrite the report with subtle yet significant changes that could impact the interpretation of the findings.

Definitions:
- descriptor error: An error in the description or characterization of an observation, finding or anatomical structure in a radiology report, without changing the observation itself or its anatomical location.
- observation: A specific finding or abnormality noted in a radiology report. An observation typically describes a particular condition, anomaly, or feature identified during the imaging study.

Please adhere to the following guidelines:
1. Introduce 1-3 descriptor errors, which can involve changing one or more of the following aspects of an observation:
   quantity, size, shape, homogeneity, density, margin, severity, temporality, distribution, enhancement, temporal change, certainty, composition, complexity, normalcy, caliber, malignancy status, patency, occupancy, integrity, direction, component involved, or position.
2. Avoid making changes to the observations themselves, adding observations (observation errors), or changing their anatomical locations (anatomical errors).
3. Ensure the errors are plausible and maintain medical coherence within the context of the report. You can modify the rest of the report to be consistent with that error. The report should still make sense medically, even with the introduced errors.
4. Preserve the overall structure and style of the original report, even in the errors introduced.

Please provide the modified report with introduced descriptor errors while maintaining overall medical coherence. Additionally, include information about which errors were introduced.

Your response should be in JSON format with two main keys: "modified_report" for the rewritten report, and "errors" for describing the introduced errors.
Original Radiology Report:
{report}

Please provide the rewritten report with 1-3 descriptor errors in JSON format. Use the following structure:

{
    "modified_report": "Your rewritten report here",
    "errors": [
        {
            "type": "error type: [descriptor type]",
            "description": "brief description of the error"
        },
        ...
    ]
}
```

Appendix C. Prompt for NER Annotation

```
You are a radiologist performing clinical term extraction from the FINDINGS and IMPRESSION sections in the given radiology report.
Here a clinical term can be in ['anatomy','observation_present','observation_absent','device_present','device_absent','procedure','descriptor'].
'anatomy' refers to the anatomical body, such as 'left frontal scalp', 'paranasal sinus';
'observation_present' refers to findings, diseases are present according to the sentence, such as 'haemorrhage', 'lesion';
'observation_absent' refers to findings, diseases or medical devices are not present according to the sentence;
'device_present' refers to medical devices are present according to the sentence, such as 'ventricular drain', 'stent', 'clip';
'device_absent' refers to medical devices are not present according to the sentence;
'procedure' refers to procedures are used to diagnose, measure, monitor or treat problems;
'descriptor' refers to modifiers used to describe the observation, including categories as quantity ('single', 'isolated', 'multiple',etc), size ('large','big',etc), shape ('lobulated','round',etc), homogeneity('homogeneous','heterogeneous',etc), density ('isodense','hypodense',etc), margin ('well-defined','poor-defined',etc), severity ('mild','moderate','severe', etc), temporality ('acute','new','old','age-indeterminate',etc), destribution ('confluent','diffuse',etc), enhancement, change ('decreased','worsened', etc), certainty ('probably','like'), composition ('gas','fluid','solid','mixed',etc), complexity ('simple','complex'), normalcy ('normal','abnormal'), caliber('detailed','normal'), malignancy_status ('benigh','malignant'), patency ('patent','occlusion'), occupancy ('empty'.'full'), integrity ('intact','partially ruptured').

For example, the sentence 'Similar 3.6 x 2.7 cm left thalamus dense hematoma with surrounding low-density vasogenic edema.' '3.6 x 2.7 cm' is 'descriptor', 'left thalamus' is 'anatomy', 'dense' is 'descriptor', 'hematoma' is 'observation_present', 'low-density' is 'descriptor', 'vasogenic edema' is 'observation_present'.
Note that for the sentence 'Gray-white matter differentiation is within normal limits.', 'Gray-white matter differentiation' should be 'observation_absent', and 'within normal limits' should be 'descriptor'.
Note that for sentences include 'loss of gray-white differentiation', should be 'observation_present'. Note that for 'CT','MRI','contrast administration','post-contract','clinical follow-up' are not procedure. for sentences that do not describe specific patient imaging findings but instead comment on the general capabilities or limitations of noncontrast CT, such as 'Acute infarct may be inapparent on noncontrast CT.'. return an empty object {}.
Note that observation should not contain the anatomy and descriptor.
Given a sentence, and reply with the JSON format following template: {'<sentence>':{'entity':'entity type','entity':'entity type'}}
```

```
You are a radiologist performing entity relation extraction from the FINDINGS and IMPRESSION sections in the given radiology report. Here a entity can be in ['anatomy','observation_present','observation_absent','device_present','device_absent','procedure','descriptor'].
The relation can be in ['modify','located_at','suggestive_of','associate_with'].
The relation between entities can be  descriptor 'modify' anatomy, descriptor 'modify' observation, descriptor 'modify' device observation/device 'located_at' anatomy, observation/device 'suggestive_of' observation/device, observation/device 'associate_with' observation/device.

For example, the sentence 'Similar 3.6 x 2.7 cm left thalamus dense hematoma with surrounding low-density vasogenic edema.' '3.6 x 2.7 cm' is 'descriptor', 'left thalamus' is 'anatomy', 'dense' is 'descriptor', 'hematoma' is 'observation_present'. 'low-density' is 'descriptor', 'vasogenic edema' is 'observation_present'. It should be '3.6 x 2.7 cm' modify 'hematoma', 'dense' modify 'hematoma', 'hematoma' located_at 'left thalamus', 'low-density' modify 'vasogenic edema', 'vasogenic edema' located_at 'left thalamus'.
    'Skull: No acute calvarial abnormality.', 'Skull' is 'anatomy', 'No' is 'descriptor', 'acute' is 'descriptor', 'calvarial' is 'anatomy', 'abnormality' is 'observation_absent'.
```

```
     It should be ‘No’ modify ‘abnormality’, ‘acute’ modify ‘abnormality’, ‘abnormality’
    located_at ‘calvarial’.

Given a radiology report, with entities and their types for each sentence like {‘<sentence
    >’:{‘entity’:‘entity type’,‘entity’:‘entity type’},‘<sentence>’:{‘entity’:‘entity type
    ’,‘entity’:‘entity type’}  , determine the relationships between these entities.

Reply with the extracted relationships in the following JSON format: {‘<sentence>’:[[
    source_entity,relation,target_entity],[...]],‘<sentence>’:[[source_entity,relation,
    target_entity],[...]]}
```