Building blocks for complex tasks: Robust generative event extraction for radiology reports under domain shifts

Sitong Zhou[♠] **Meliha Yetisgen** [♢] **Mari Ostendorf**[♠]
[♠]University of Washington, Electrical & Computer Engineering
[♢]University of Washington, Biomedical and Health Informatics
{sitongz,melihay,ostendor}@uw.edu

Abstract

This paper explores methods for extracting information from radiology reports that generalize across exam modalities to reduce requirements for annotated data. We demonstrate that multi-pass T5-based text-to-text generative models exhibit better generalization across exam modalities compared to approaches that employ BERT-based task-specific classification layers. We then develop methods that reduce the inference cost of the model, making large-scale corpus processing more feasible for clinical applications. Specifically, we introduce a generative technique that decomposes complex tasks into smaller subtask blocks, which improves a single-pass model when combined with multitask training. In addition, we leverage target-domain contexts during inference to enhance domain adaptation, enabling use of smaller models. Analyses offer insights into the benefits of different cost reduction strategies.

1 Introduction

Radiology reports contain a diverse and rich set of clinical abnormalities documented by radiologists during their interpretation of the images. Automatic extraction of radiological findings would enable a wide range of secondary use applications to support diagnosis, triage, outcomes prediction, and clinical research (Lau et al., 2020). We adopt an event-based schema to capture both indications, the reason for radiology exams, and abnormal findings documented in radiology reports. We use an annotated a corpus of reports from three distinct radiology examination modalities (Lybarger et al., 2022): Magnetic Resonance Imaging (MRI), Positron Emission Tomography (PET), and Computed Tomography (CT). Each event consists of a trigger, words that indicate a particular indication or finding (e.g., lesion), and a set of attributes (assertion, anatomy, characteristics, size, size trend, size count) that describe this indication or finding. Manual annotation of radiology reports is costly, therefore we hope models can generalize across different exam modalities. In this work, we define each modality in our annotated corpus as a domain and study cross-domain generalization among different modalities for the task of event extraction. Event extraction can be conceptualized as a series of subtasks, which include entity detection (trigger and attribute spans), relation detection (between triggers and attributes), and entity normalization (fine-grained labels on spans). In our experiments, we focus on trigger detection and anatomy attribute extraction with normalized labels.

To enhance generalization capabilities, some studies employ generative models and formulate tasks as question answering and using texts to represent both inputs and outputs (Raffel et al., 2020; Xie et al., 2022), as opposed to allowing the model to solely learn task intent from training data (Eberts and Ulges, 2019; Lybarger et al., 2023).

The exceptional performance of generative models often rely on large model size; however, in real-time inference for processing large-scale clinical notes, reducing inference costs is crucial. To address this need, for task inference, we want to reduce the number of decoding passes and employ smaller models. Due to the high inference costs, there is a desire to merge these subtasks and decode them in a single step. However, the generative approach has been reported to perform better on solving subtasks individually but worsen when combined, a phenomenon referred to as the compositionality gap (Press et al., 2022). This gap can be exacerbated under domain shifts when models learn subtasks jointly, as interdependence of subtasks may vary across domains.

While large language models (LLMs) mitigate the compositionality gap using reasoning steps (Wei et al., 2022; Press et al., 2022) to solve complex questions by decomposing them into smaller ones, there is limited work on reasoning for highly specialized domains (such as medical event extrac-

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 344–357

tion) or with smaller models. In this paper, we reduce the compositionality gap for smaller models through formatting of complex tasks into easier subtasks as blocks. This approach teaches models how to solve individual subtasks independently and how to assemble them for solving more complex tasks.

The generative model enables seamless integration of supplementary contexts into the prompt, which compensates for the knowledge gap to larger models and reduces inference costs. To aid in domain adaptation, we extract target domain contexts that are likely to be helpful for the task, instead of retrieving similar contexts for general purpose. Specifically, to assist with anatomy normalization tasks, we employ an unsupervised extractor to acquire pertinent contexts that likely contain anatomical information from the same document and/or unannotated text from the same domain. This process can either disambiguate the original single-sentence input or provide anatomy-related hints that the model can utilize. To avoid introducing source-domain-specific reliance on the contexts, we incorporate the contexts only at the inference stage.

In our experiments, we first study domain shift for extracting radiology finding events and observe that cross-domain performance decline is more pronounced for knowledge-intensive anatomy normalization tasks, while detecting entity spans exhibits relatively stable performance. We demonstrate that building subtask blocks and assembling them as sequences to solve complex tasks can reduce the compositionality gap in smaller models. We show that incorporating target-domain contexts in domain adaptation can compensate for reduced model sizes, enabling good performance with smaller models.

2 Task

2.1 Event extraction for radiology findings

Our event scheme includes three event types: i) *Indication* is the reason for the imaging (e.g. motor vehicle accident or cancer staging); ii) *Lesion* captures lesions uncovered by the exam (e.g. mass or tumor); and iii) *Medical Problem* characterizes non-lesion abnormalities (e.g. fracture or hernia). Each finding event is characterized by an event trigger and set of attributes (assertion, anatomy, characteristics, size, size-trend, count). In this work, we focus only on extracting events with normalized anatomical information and investigate cross-domain generalization for different examination modalities. Figure 1 presents a *Lesion* event example. The event extraction process can be broken down into four subtasks: (1) Trigger span extraction (e.g., "density"), (2) Trigger type classification (e.g., "density" - Lesion), (3) Anatomy span extraction (e.g., "left lobe of liver" associated with the trigger "density"), and (4) Anatomy normalization to parent-child anatomy categories (e.g., "left lobe of liver" - Parent: Hepato-Biliary, Child - Liver). See Appendix A for the full list of hierarchical parent-child anatomy categories.

We evaluate event extraction performance using the F1 metrics by Lybarger et al. (2021). Our assessment of the trigger extraction is based on the span overlap and the event type match with respect to the gold standard labels. The anatomy extraction is first assessed at the span level. A correct anatomy prediction is associated with a correct predicted trigger and anatomy span overlap with the gold standard labels. Additionally, we evaluate anatomy extraction based on the normalization level, irrespective of their spans. A match between the predicted anatomy entity and the gold label indicates that the trigger is matched, and the normalized anatomy category is equal.

2.2 Domain shifts across radiology modalities

Our research investigates cross-domain generalization among three distinct radiology examination modalities: MRI, PET, and CT. These exam modalities are performed for different reasons with different technologies and the resulting radiology reports differ in terms of level of details as well as anatomy distribution. While CT and MRI scans allow radiologists to view structures inside the body, a PET scan, on the other hand, captures how tissues in the body work on the cellular level and shows unusual activity. MRI scans very frequently involve neurological exams. The most common use of PET scans is to diagnose or monitor certain cancer types. In our experiments, we define each modality as a domain. We use PET as the target domain, and train on three domains separately to evaluate both in-domain and cross-domain scenarios.

3 Method

3.1 Generative event extraction with T5

In order to improve the model's generalization capabilities over BERT-based alternatives (Lybarger et al., 2023; Eberts and Ulges, 2019), we struc-

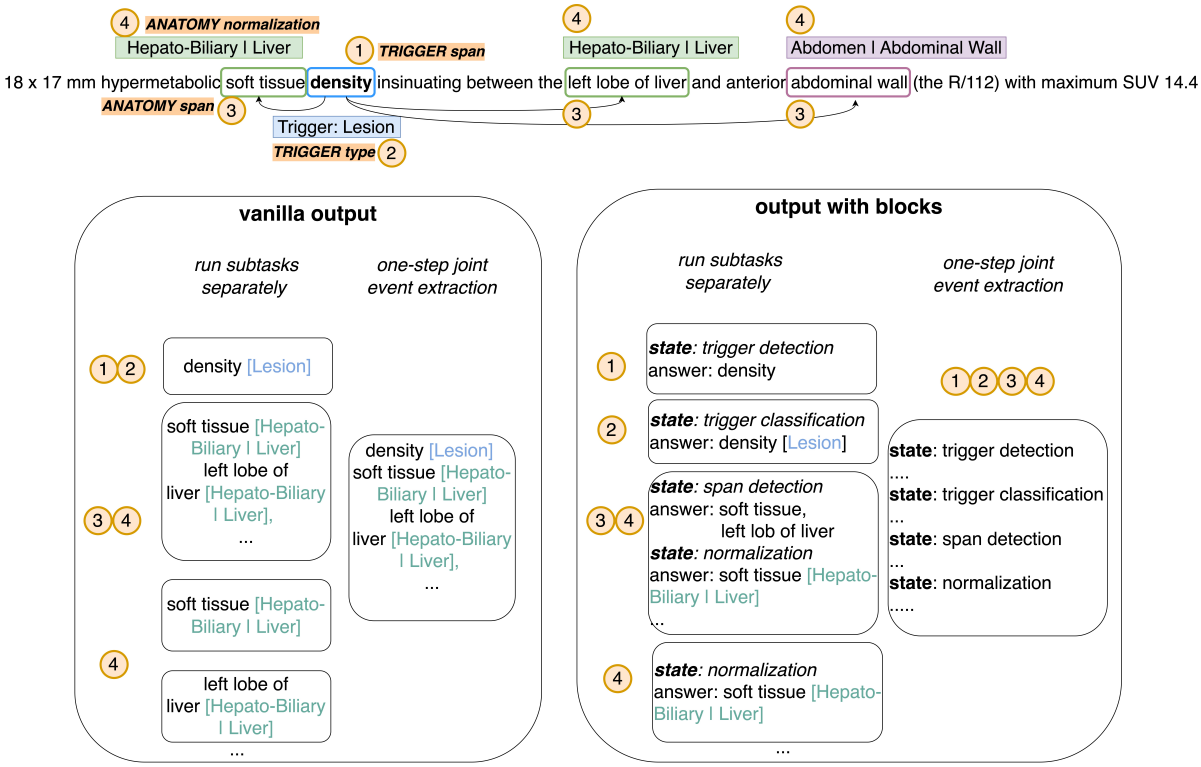

Figure 1: Representations of anatomical information in radiology reports, with the event-based annotation at the top and two generative model output formats to multi-step and one-step processing. The left-hand side shows the **vanilla format** and the right-hand side shows the **building block format**.

ture our event extraction task in a unified question-answering (QA) format (Xie et al., 2022; Raffel et al., 2020). With the generative approach, the model leverages the semantic meaning of prompts for specifying subtasks and associated categorization labels. Based on experiments with in-context learning (Hu et al., 2022), we expect this to be beneficial for domain-mismatches in class label distributions, e.g. where infrequent classes in the source domain are frequent in the target domain. Furthermore, the text-to-text format offers the flexibility to incorporate additional contexts to facilitate tasks, as discussed in Section 3.3.

The input prompt comprises: (1) an input sentence from clinical notes to extract events from, (2) a question that describes the task or subtask, and (3) an ontology that provides textual labels for classification tasks and hierarchical relationships if multi-level granularities are required. The output is a word sequence that specifies the extracted information (the answer). Two alternative output formats are discussed in the next section; example input-output pairs for both are in Appendix B.

Event extraction can be seen as a multi-hop question-answering process, involving a series of subtasks for successful completion. We use a pipeline approach to address the event extraction subtasks in different steps, where each step in the pipeline consists of a specialized generative model trained for one or more of the subtask types. Three different architectures are explored:

Three-step approach: This involves a first step for detecting trigger spans and trigger types, followed by a second step for identifying the anatomy associated with each detected trigger, and a third step for normalizing each identified anatomical entity at parent and child levels individually.

Two-step approach: This involves a first step for detecting trigger spans and trigger types, followed by a second step for identifying and normalizing the anatomy associated with each detected trigger.[1]

One-step approach: we address all subtasks, which may be associated with multiple entities, in a single pass per input sentence. This method results in longer output lengths compared to the individual steps of previous two approaches.

The one-step approach substantially reduces in-

[1]Both the 2-step and 3-step approaches use the same second step, predicting anatomy spans and their normalized values. The three-step approach drops the normalized values from its second step.

ference costs compared to other two multi-step approaches. However, we find that it negatively impacts model performance due to the longer output and the compositionality gap. The performance loss is mostly recovered by changing the output format (as described next) together with a multi-task training strategy. Specifically, we train the model on both the complete task and the decomposed subtasks. This allows the model to perform subtasks independently and assemble subtask sequences for complex tasks. During inference, we decode in a single step to minimize costs.

Our work builds on generative models, specifically the clinical version of the T5 models (Lu et al., 2022), which are pre-trained on medical articles and clinical notes. This choice leverages their strengths in comprehending clinical text styles and medical knowledge.

3.2 Output formats

We explore two different output formats as illustrated in Figure 1, with subtask answers provided in sequence when there are multiple subtasks.

The baselines leverage a standard output format (referred to here as the **vanilla format**), which specifies the answer for a subtask with an extracted span followed by the entity label in brackets "[]". When multiple entities are detected, they are generated in sequence.

The vanilla format can be used with the one-step approach, but the resulting output can be very long when multiple triggers and/or entities are detected. The lack of distinction between types of spans in the output makes it harder for the language model to learn the subtask structure. To address this problem, we introduce a state-augmented prompt (referred to as the **building block** format), in which each subtask is associated with a state (as in a finite-state transducer) and explicitly named. Our approach is motivated by the work on chain-of-thought LLMs (Wei et al., 2022; Press et al., 2022), which use natural language reasoning in the generated outputs to address the compositionality gap. However, it differs in that we do not use natural language reasoning, but rather more of a programming-like description. In addition, the finite-state framework is amenable to meulti-task training, which is particularly important for the block approach.

3.3 Using target-domain contexts in prompts

A single input sentence may not provide enough information for a model to complete a task, as additional details may be needed for disambiguation or to supplement missing knowledge in pre-trained language models. Fortunately, the text format of the input allows for the seamless integration of additional contexts from the target domain during inference to aid in the task and infuse helpful domain-specific bias, even if the models were not trained for reading contexts.

The desired contexts should be relevant to the input sentence and contain helpful task information. We utilize two types of contexts: document-level and domain-level contexts to help anatomy normalization subtasks. Document-level contexts include adjacent sentences before and after the input, automatically extracted section headers[2] and exam type metadata associated with the same clinical note. The document-level contexts are likely to describe relevant anatomical parts, as section headers and exam types often summarize anatomical information. Domain-level contexts are retrieved from the unlabeled target-domain corpus. We search for the most similar sentence with the greatest lexical overlap degree, using the search algorithm BM25 (Trotman et al., 2014).[3] When the search pool is large, the top-ranked retrieved context sentence likely describes a similar anatomy part as the queried input sentence. To reduce computational costs and ensure that the retrieved sentences contain useful anatomical information, we pre-filter the target corpus to limit the search scope to sentences containing common anatomy terms listed from anatomy normalization categories and high-frequency auto-extracted section headers, reducing the number by 74%. More context-retrieval details are in Appendix D.

We add contexts only during decoding (and not in training) to prevent the model from relying too much on source-domain contexts. In the input prompts, exam type, section headers and prior sentences are placed before input sentences, following their natural orders. Other contexts are inserted between the input sentences and task ontology.[4] We test this approach in a separate anatomy nor-

[2]We extract section headers as the beginning of the last previous sentence containing ':'

[3]We implement the BM25 algorithm using https://github.com/dorianbrown/rank_bm25

[4]The full T5 input template is described in Table 9 from Appendix B

malization run after the one-step building block model. This process combines building block output format with target domain context integration. The reason for not directly adding it to a one-step process is that introducing contexts to inputs can potentially corrupt span detection, as the model may extract spans from the context rather than exclusively from the input sentence.

4 Experiments

4.1 Radiology datasets across exam modalities

Data split	Note Count	Sent Count
CT (train)	143	3707
MRI (train)	144	3551
PET (train)	142	5184
PET (valid)	20	758
PET (test)	40	1481
PET (unlabeled)	1471	50000

Table 1: Dataset statistics for the three radiology examination modalities: CT, MRI, and PET. We explore in-domain and cross-domain training, evaluating on PET.

We use an annotated corpus containing radiology notes about CT, MRI, and PET imaging exams; statistics are given in Table 1. The anatomy normalization labels are grouped into sublevels according to the SNOMED CT concepts. Notes in the test and validation sets are all doubly annotated. The inter-rater agreement for Trigger is 0.73 F1.

Variations in anatomy distribution across imaging modalities can cause domain discrepancies. PET has the most balanced distribution among parent-level anatomy categories, followed by CT. However, MRI has a heavily skewed distribution, with 62% of trigger-associated anatomy entities being neurological among 16 parent-level categories. See Appendix A for anatomy distribution details.

To enhance domain-specific context retrieval and boost the chances of retrieving helpful contexts, we expand the search pool by sampling 50,000 unlabeled PET report sentences from the same distribution as in the annotated reports (Lybarger et al., 2022), with a minimum of three tokens.

4.2 Implementation

In the non-generative baseline, we adopt the mSpERT model (Lybarger et al., 2023) for hierarchical multi-label entity and relation extraction. Entities are extracted as spans. We initialize with Bio-Clinical BERT (Alsentzer et al., 2019).

For the T5 model using both vanilla output formats and the subtask block formats, we initialize with ClinicalT5 (Lu et al., 2022),

For all models, the best checkpoint is chosen after 15 training epochs based on the validation performance on the target domain. For T5 models with multitask training on subtask blocks, which involves a higher number of training steps, we evaluate the model on the validation set after every 0.5 epoch approximately. For methods that do not involve multitask training, we evaluate the model on the validation set per epoch.

We implement multitask training on subtask blocks for MRI and PET, using the auxiliary tasks, as described in Section 3.1, including trigger span detection, trigger classification, joint anatomy span detection and normalization, and anatomy normalization. For the CT-PET transfer scenario, we add an additional anatomy span detection auxiliary task, as we observe that more aggressive learning is needed for anatomy span detection in the CT domain. Detailed information about hyperparameters can be found in Appendix E.

5 Results

Table 2 shows the trigger and anatomy detection results for mSpERT compared to different context-independent T5-base alternatives. For the in-domain condition, all T5 approaches outperform the mSpERT model for the three anatomy-related metrics. The results for trigger detection are mixed, but fairly similar for all. The best performance overall is obtained using the 2-step vanilla output T5 model. For the cross-domain scenarios, all models suffer degradation in performance compared to the in-domain condition, with the greatest performance drop for the normalized anatomy categories, particularly for the MRI-PET condition which has the greatest mismatch in anatomy distribution. The performance loss is greatest for the mSpERT model, with a 44% relative reduction in F1 scores for normalized anatomy (at both parent and child levels) for the MRI-PET case. In contrast, the relative loss on the parent and child levels for the T5 models is 24-29%. For both within and across-domain scenarios, the building block technique improves the 1-step results for all categories, but particularly for the more difficult anatomy normalization tasks. As described later in Section 6.2, the 1-step

Table 2: F1 scores (%) for: non-generative mSpERT (Lybarger et al., 2023), generative vanilla T5 models with both multi-step pipeline and one-step joint approaches, and our proposed one-step T5 model using the building block technique. All models adopt the T5-base architecture and are initialized with ClinicalT5 (Lu et al., 2022). **Best overall** scores are in bold, and best one-step scores are underlined.

Entity	mSpERT	T5-base 3-step (vanilla)	T5-base 2-step (vanilla)	T5-base 1-step (vanilla)	T5-base 1-step (blocks)
			PET → PET		
Trigger	82.4	81.9	81.9	82.1	**82.6**
Anatomy Span	65.8	**67.6**	**67.6**	66.0	66.1
Anatomy Parent	61.9	64.7	**64.9**	63.3	63.5
Anatomy Child	59.6	62.1	**62.3**	59.7	60.7
			MRI → PET		
Trigger	75.6	76.6	76.6	76.4	**77.8**
Anatomy Span	59.9	60.9	60.9	59.2	**61.1**
Anatomy Parent	34.7	**48.6**	47.1	44.9	48.3
Anatomy Child	33.5	44.6	44.0	41.2	**44.8**
			CT → PET		
Trigger	75.7	76.1	76.1	74.0	**76.6**
Anatomy Span	59.7	**61.4**	**61.4**	56.3	59.8
Anatomy Parent	53.2	**55.8**	54.8	50.8	55.0
Anatomy Child	47.5	**53.3**	51.8	48.1	51.2

approach is sensitive to the compositionality gap, which is ameliorated by the block approach. For the cross-domain scenarios, the best overall results are obtained with the 3-step approach for the CT-PET condition and with the 1-step block approach for the MRI-PET condition (greater mismatch). An additional advantage of the 1-step approach is the lower latency associated with using only one decoding pass.

As described earlier, target-domain contexts are added to prompts during a second step of T5 decoding to help anatomy normalization, after the 1-step subtask block decoding with T5-base. Table 3 shows results for all different types of contexts, as well as using either T5-large or T5-base in the second step without context. Without context, the T5-base and T5-large models give similar results for in-domain and CT-PET cross-domain conditions, but T5-large improves results for the MRI-PET condition. (Note that T5-large is only used in the last step; a bigger benefit could be observed if used in both steps.) All types of context are useful for the two domain-shift cases, but there is little or no benefit for the in-domain case. Of the different types of context, automatically retrieved similar sentences from unlabeled target-domain data provide the greatest benefit in the mismatched scenarios. Combining all contexts provides a small additional benefit, except for the anatomy parent in the MRI-PET case. Anecdotally, we observe that same-document contexts are useful for disambiguation, while hints for challenging examples are more likely collected from a large domain-level corpus rather than just the same document. (For examples, see Appendix F.)

Table 4 provides information on the relative cost of the different T5 models. The multi-pass models have higher latency (average passes/sample) in that passes are necessarily sequential. (Note that samples with no findings or no anatomy identified in the first pass do not require additional passes.) The number of tokens per sample is an indicator of cost. The 1-step model with blocks has a higher cost than the 2-step approach because of the additional tokens introduced by the state-augmented prompt, but the cost is still lower than the 3-step approach. The use of context adds additional cost.

Table 3: F1 scores (%) for T5 anatomy classification models with and without contexts. Results with context involve a first pass with the 1-step T5-base building blocks method, the same as "T5-base one-step (blocks)" in Table 2, followed by another pass that normalizes the anatomy spans that are previously detected by the 1-step T5-base (block) model. We normalize with the model used in the last step of the 3-step (vanilla) pipeline, optionally augmented with contexts in the prompts. We also add the T5-large normalization model without context to compare with the larger-scale counterpart.

Normalization model	**T5-large**	**T5-base**	**T5-base**	**T5-base**	**T5-base**	**T5-base**
Context	**n/a**	**n/a**	**adjacent sentences**	**metadata & header**	**BM25 retrieval**	**all combined**
	PET → PET, Trigger: 82.6, Anatomy Span: 66.1					
Anatomy Parent	63.6	**63.9**	63.8	63.7	63.8	63.7
Anatomy Child	60.9	60.9	61.0	**61.1**	60.3	60.4
	MRI → PET, Trigger: 77.8, Anatomy Span: 61.1					
Anatomy Parent	51.2	50.8	52.1	51.6	**53.8**	53.5
Anatomy Child	48.6	45.4	47.1	46.6	48.3	**48.8**
	CT → PET, Trigger: 77.8, Anatomy Span: 59.8					
Anatomy Parent	54.1	54.2	55.5	55.0	55.5	**55.9**
Anatomy Child	51.2	51.2	52.2	51.6	52.6	**53.0**

Table 4: Average number of decoding passes per sample (indicating relative decoding time) and tokens per sample (indicating relative cost) of one-step and multi-step approaches for testing on the PET domain. The token counts per sample are the average of the sum of input and output token counts, which is used for proportionality pricing LLM usage by ChatGPT. The context method uses all context combined in another normalization step as in Table 3.

Method	**passes/ sample**	**tokens/ sample**
3-step (vanilla)	2.5	355
2-step (vanilla)	1.7	199
1-step (block) + context	1.7	450
1-step (block)	1	245

Table 5: F1 scores (%) for the cross-domain MRI-PET condition using 1-step T5-base models, comparing: vanilla output format, building block format but no multitask training, and building block format with multitask training.

Entity	**vanilla**	**blocks, no multitask**	**blocks, multitask**
Trigger	76.4	76.0	**77.8**
Anatomy	59.2	57.1	**61.1**
Parent	44.9	38.6	**48.3**
Child	41.2	36.9	**44.8**

6 Analysis

In this section, we analyze results to better understand performance improvements associated with the subtask block format and retrieved context in prompts.

6.1 Multitask training for subtask blocks

To understand the contributing factors for the subtask block method's effectiveness, we examine whether the output format encodes helpful structural task information, or multitask training on individual subtasks predominantly drives performance. We conduct an additional experiment using the same subtask block output format, but without the multitask training for individual blocks. We use MRI as the source domain, because it suffers the most cross-domain performance drop. The results in Table 5 show a substantial drop in the model's performance in the absence of multi-task training, as compared to both the multi-task version and the baseline output format. This performance degradation may be attributed to increased decoding lengths.

6.2 Predictions for multiple anatomy parents

In addition to differences in the anatomy parent class distribution across domains, the three examination modalities also differ in how frequently sentences with multiple anatomy entities involve multiple parent classes. As shown in Table 6, 57% of the sentences with multiple anatomy entities in the target domain (PET) have multiple parents, whereas the percentage is much lower for the other domains (only 12% for MRI). When using the vanilla method, models trained on a domain with few instances of multiple parents will tend to predict the same parent class for each entity, as shown by the lower frequency of prediction in the table. The use of subtask blocks together with multitask training substantially improves the model's ability to identify multiple parent types when there are multiple anatomy entities. In all domains, roughly 20% of sentences have multiple anatomy entities, so this leads to overall performance improvement.

Table 6: Relative frequency (%) of sentences with multiple anatomy entities that have different parents, comparing frequencies as predicted by different models to the frequencies based on gold annotations for training data. The gold relative frequency on the PET test data is 55%.

Domain	Training	Vanilla	Blocks
PET	57	53	**56**
MRI	12	29	**46**
CT	33	45	**52**

6.3 Target domain retrieval filtering

Table 7: Normalized anatomy F1 score (%) for the MRI-PET condition, comparing approaches for using target-domain context retrieved using BM25: no context, unfiltered retrieval, and filtering the retrieval corpus to anatomy informative sentences.

Entity	no context	unfiltered contexts	filtered contexts
Parent	50.8	52.7	**53.8**
Child	45.4	47.4	**48.3**
Trigger: 77.8, Anatomy: 61.1			

To reduce the search costs, we filter the unlabeled target domain data to include only sentences with anatomy terms before running retrieval with BM25. To understand the impacts on performance, we run experiments on unfiltered data, again focusing on the MRI data where domain differences are greatest. Table 7 shows that filtering for anatomy not only reduces costs but also gives a small improvement in results for identifying normalized categories.

7 Related work

7.1 Event extraction methods

Event extraction research has predominantly depended on BERT-based (Devlin et al., 2019; Alsentzer et al., 2019) models, where the extraction subtasks are performed by classifiers utilizing the language model layer representations (Eberts and Ulges, 2019; Zhong and Chen, 2021; Lybarger et al., 2023). They often yield satisfactory results when training on sufficient in-domain training data. For example, when training and testing on CT scan reports, normalizing anatomical terms can result in an F1 score of 79% for nine major body parts and 73% for 41 sub-body parts (Lybarger et al., 2021). Recently, there has been growing interest in adopting generative approaches (Raffel et al., 2020; Brown et al., 2020) for information extraction, which incorporates task descriptions and auxiliary context information to enhance performance (Xie et al., 2022). Many efforts (Lu et al., 2022; Phan et al., 2021; Lehman et al., 2023; Luo et al., 2022) support exploration of clinical tasks through pre-training generative models for biomedical and clinical domains. In this study, we explicitly evaluate generative models in domain shift settings, with an emphasis on minimizing inference costs.

7.2 Context augmentation

Integrating models with supplementary contexts has shown benefits in knowledge-intensive tasks (Lewis et al., 2020; Guu et al., 2020). Generative models can utilize knowledge prompts from external knowledge sources (Peng et al., 2023; Liu et al., 2021). In our work, we retrieve contexts from the unlabeled clinical note corpus without relying on external resources.

7.3 Compositionality Gap

The compositionality gap has been identified as a challenge in generative models when multiple subtasks are combined (Press et al., 2022). Prior research on large language models has demonstrated that breaking down complex tasks into smaller subproblems can be beneficial (Wei et al., 2022; Press

et al., 2022). Small models have been employed for multiple decoding passes (Khot et al., 2021), but there is limited research on reasoning with smaller models that merge these steps, which is essential for real-time applications in the clinical field.

8 Conclusion

In conclusion, we present generative event extraction methods for radiology findings that improve generalization under domain shifts and reduce the inference costs. By decomposing complex tasks into simpler subtask blocks and incorporating target-domain context during the inference process, our approach enables smaller models to achieve performance similar to or better than those obtained with more decoding passes, and comparable to larger models on anatomy normalization. Our methods make efficient inference for extensive clinical notes more feasible. This work offers insights into reasoning with smaller models and using context to compensate the reduced model size.

Limitations

The use of machine learning models in clinical decision-making requires an understanding of the reasoning behind model predictions. Our study focuses on improving the performance of smaller models using context and subtask blocks. While the subtask state labels provide some interpretability, we have not explored its impact on trust among medical professionals. In addition, the relative benefit of the different multi-pass strategies and different types of context appear to depend on the degree of domain mismatch, which should be further explored in future work.

Ethics Statement

Radiology reports contain sensitive patient information and it is crucial to handle this data responsibly, adhering to strict privacy and confidentiality guidelines. The dataset used in this paper was fully-de-identified. We received approval from our institution's IRB prior to conduct the presented research and used HIPAA compliant servers. Additionally, a careful examination is needed to assess potential bias in models used for extracting information from radiology reports prior to implementing real life secondary use applications.

Acknowledgements

This work was supported by NIH/NCI (1R01CA248422-01A1 and 1R21CA258242-01).

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Conference on Neural Information Processing Systems*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Markus Eberts and Adrian Ulges. 2019. Span-based joint entity and relation extraction with transformer pre-training. In *European Conference on Artificial Intelligence*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. REALM: Retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Yushi Hu, Chia-Hsuan Lee, Tianbao Xie, Tao Yu, Noah Smith, and Mari Ostendorf. 2022. In-context learning for few-shot dialogue state tracking. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2627—-2643. Association for Computational Linguistics.

Tushar Khot, Daniel Khashabi, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2021. Text modular networks: Learning to decompose tasks in the language of existing models. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, page 1264–1279.

Wilson Lau, Laura Aaltonen, Martin L. Gunn, and Meliha Yetisgen-Yildiz. 2020. Automatic assignment of radiology examination protocols using pretrained language models with knowledge distillation. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2021:668–676.

Eric P. Lehman, Evan Hernandez, Diwakar Mahajan, Jonas Wulff, Micah J. Smith, Zachary M. Ziegler, Daniel Nadler, Peter Szolovits, Alistair E. W. Johnson, and Emily Alsentzer. 2023. Do we still need clinical language models? *ArXiv*, abs/2302.08091.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. 2021. Generated knowledge prompting for commonsense reasoning. In *Annual Meeting of the Association for Computational Linguistics*.

Qiuhao Lu, Dejing Dou, and Thien Nguyen. 2022. ClinicalT5: A generative language model for clinical text. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5436–5443, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. BioGPT: Generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*.

Kevin Lybarger, Aashka Damani, Martin L. Gunn, Özlem Uzuner, and Meliha Yetisgen-Yildiz. 2021. Extracting radiological findings with normalized anatomical information using a span-based BERT relation extraction model. *AMIA ... Annual Symposium proceedings. AMIA Symposium*, 2022:339–348.

Kevin Lybarger, Nicholas J Dobbins, Ritche Long, Angad Singh, Patrick Wedgeworth, Özlem Uzuner, and Meliha Yetisgen. 2023. Leveraging natural language processing to augment structured social determinants of health data in the electronic health record. *Journal of the American Medical Informatics Association*. Ocad073.

Kevin Lybarger, Namu Park, Sitong Zhou, Aashka Damani, Alison Brennan, Jagjeet Gill, Nianiella Dorvall, Vy Huynh, Spencer Lewis, Martin L. Gunn, Özlem Uzuner, and Meliha Yetisgen-Yildiz. 2022. A corpus of radiology reports from multiple imaging modalities with fine-grained event-based annotations. In *American Medical Informatics Association Annual Symposium*.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Lidén, Zhou Yu, Weizhu Chen, and Jianfeng Gao. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *ArXiv*, abs/2302.12813.

Long Phan, James T. Anibal, Hieu Trung Tran, Shaurya Chanana, Erol Bahadroglu, Alec Peltekian, and Grégoire Altan-Bonnet. 2021. SciFive: a text-to-text transformer model for biomedical literature. *ArXiv*, abs/2106.03598.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. 2022. Measuring and narrowing the compositionality gap in language models. *ArXiv*, abs/2210.03350.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Andrew Trotman, Antti Puurula, and Blake Burgess. 2014. Improvements to BM25 and language models examined. *Proceedings of the 19th Australasian Document Computing Symposium*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Tianbao Xie, Chen Henry Wu, Peng Shi, Ruiqi Zhong, Torsten Scholak, Michihiro Yasunaga, Chien-Sheng Wu, Ming Zhong, Pengcheng Yin, Sida I. Wang, Victor Zhong, Bailin Wang, Chengzu Li, Connor Boyle, Ansong Ni, Ziyu Yao, Dragomir R. Radev, Caiming Xiong, Lingpeng Kong, Rui Zhang, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. 2022. UnifiedSKG: Unifying and multi-tasking structured knowledge grounding with text-to-text language models. In *Conference on Empirical Methods in Natural Language Processing*.

Zexuan Zhong and Danqi Chen. 2021. 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, Online. Association for Computational Linguistics.

A Hierarchical anatomy normalization categories

We normalize detected anatomy spans for applications focusing on anatomy categories rather than specific anatomy terms. We classify at different granularities, a parent-level coarse classification with 16 parent labels and a child-level fine-grained

Parent-level Class	Child-level Classes
Neurological	Undetermined, Spine Cervical, Spine Thoracic, Spine Lumbar, Spine Sacral, Spine Cord, Spine Unspecified, Brain, Nerve, Pituitary, Cerebrospinal Fluid Pathway, Cerebrovascular System, Extraaxial
Cardiovascular	Undetermined, Venous, Arterial, Pulmonary Artery, Heart, Pericardial Sac, Coronary Artery
Thoracic	Undetermined, Mediastinal
Respiratory	Undetermined, Lung, Pleural Membrane, Tracheobronchial
Digestive	Undetermined, Esophagus, Stomach, Intestine, Small Intestine, Large Intestine
Hepato-Biliary	Undetermined, Gallblader, Bile Duct, Pancreas, Liver
Urinary	Undetermined, Kidney, Urinary Bladder, Ureter
Lymphatic	Undetermined
F Reproductive Obstetric	Undetermined, Breast, Ovary, Uterus, Adnexal, Extra-embryonic, Placenta, Fetus, Umbilical Cord, Female Genital Structure
M Reproductive	Undetermined, Prostate, Testis, Epididymis
Musculo-Skeletal	Undetermined, Skeletal and or Smooth Muscle, Bone and or Joint
Body Regions	Undetermined, Entire Body, Pelvis, Lower Limb, Upper Limb
Head Neck	Undetermined, Thyroid, Neck, Ear, Eye, Mouth, Nasal Sinus, Pharynx, Laryngeal
Skin	Undetermined, Skin and or Mucous Membrane, Subcutaneous
Abdomen	Undetermined, Retroperitoneal, Abdominal Wall, Peritoneal Sac, Spleen, Adrenal Gland, Mesentery
Miscellaneous	Undetermined, Adipose Tissue, Connective Tissue, Biomedical Device

Table 8: Hierarchical anatomy normalization categories at parent and child levels.

classification with 72 categories. Each parent-level class includes an "Undetermined" child-level class to account for cases that don't fit into its other specified child classes. The full normalization categories are in Tabel 8.

As shown in Figure 2, MRI exhibits a more imbalanced distribution, with a majority of the anatomies related to the "Neurological" parent-level class. In CT exams, "Respiratory" account for 16% and "Neurological" represent 19% among all finding-related anatomies. For MRI, "Musculo-Skeletal" constitutes 18% while "Neurological" exams make up a substantial 62%. Lastly, in PET, "Head Neck" accounts for 12% and "Musculo-Skeletal" comprises 14%.

B Generative method input and output formats

We document the templates for the input and output, with examples in Tabel 9. For the template with contexts, "prepended contexts" include prior sentences, section headers, exam type metadata, other contexts are "appended contexts".

TEMPLATE
Template: [Input sentence] [Question] structured knowledge: [Task ontology] **Template with contexts**: [Prepended contexts] [Input sentence] [Question] structured knowledge: [Appended contexts] [Task ontology]
Trigger task ontology (for T5-vanilla pipeline: 1st step) Indication | Lesion | Medical_Problem **Anatomy task ontology** (for T5-vanilla pipeline: 2nd step, 3rd step) Neurological: Undetermined, Spine Cervical, Spine Thoracic .... (see Table A) **Trigger anatomy task ontology** (for all related to one-step building block) trigger types: Indication | Lesion | Medical_Problem anatomy categories: Neurological: Undetermined, Spine, ...
EXAMPLE
Input sentence: 18 x 17 mm hypermetabolic soft tissue density insinuating between the left lobe of the liver and anterior abdominal wall (the R/112) with maximum SUV 14.4 .
Model: T5-vanilla pipeline: first step (trigger span & type) **Question**: Question: What are medical findings in this sentence? **Output**: trigger: density [Lesion]
Model: second step (anatomy span & type) **Question**: Consider the medical finding "density" in the span "hypermetabolic soft tissue density insinuating between the", Question: What anatomy it occurs in? Where is it located? **Output**: anatomies: soft tissue [Hepato-Biliary | Liver], left lobe of the liver [Hepato-Biliary | Liver], anterior abdominal wall [Abdomen | Abdominal Wall]
Model: third step (anatomy normalization) **Question**: Consider the anatomy "soft tissue" in the span "17 mm hypermetabolic soft tissue density insinuating between", which anatomy category it belongs to among listed options? **Output**: anatomies: soft tissue [Hepato-Biliary | Liver]
Model: T5-vanilla one-step (trigger span & type, anatomy span & normalization) **Question**: Question: What are medical findings in this sentence? What anatomy they occur in? which anatomy category they belong to among listed options? **Output**: trigger: density [Lesion] anatomies: soft tissue [Hepato-Biliary | Liver], left lobe of the liver [Hepato-Biliary | Liver], anterior abdominal wall [Abdomen | Abdominal Wall]
Model: T5 one-step subtask blocks (trigger span & type, anatomy span & normalization) **Question**: [same as T5-vanilla one-step] **Output**: state: trigger detection answer: density state: trigger classification answer: density [Lesion] state: span detection answer: soft tissue, left lobe of the liver, anterior abdominal wall state: classification answer: soft tissue [Hepato-Biliary | Liver] state: classification answer: left lobe of the liver [Hepato-Biliary | Liver] state: classification answer: anterior abdominal wall [Abdomen | Abdominal Wall]
Model: multitask for trigger classification (trigger type) **Question**: Consider the medical finding "density", Question: What is the type of this medical finding? **Output**: state: trigger classification answer: density [Lesion]
Model: multitask for anatomy span (anatomy span) **Question**: Consider the medical finding "density" in the span "hypermetabolic soft tissue density insinuating between the", Question: Please identify terms that describe the finding's anatomy locations. **Output**: state: span detection answer: soft tissue, left lobe of the liver, anterior abdominal wall

Table 9: Templates and examples for T5 inputs and outputs. The "multitask" rows correspond to auxiliary tasks for the T5 one-step subtask block method. We omit rows for "multitask for anatomy" and "multitask for anatomy normalization", since they use the same question format as the 2nd and 3rd steps of the pipeline approach, but with answers in the subtask block format.

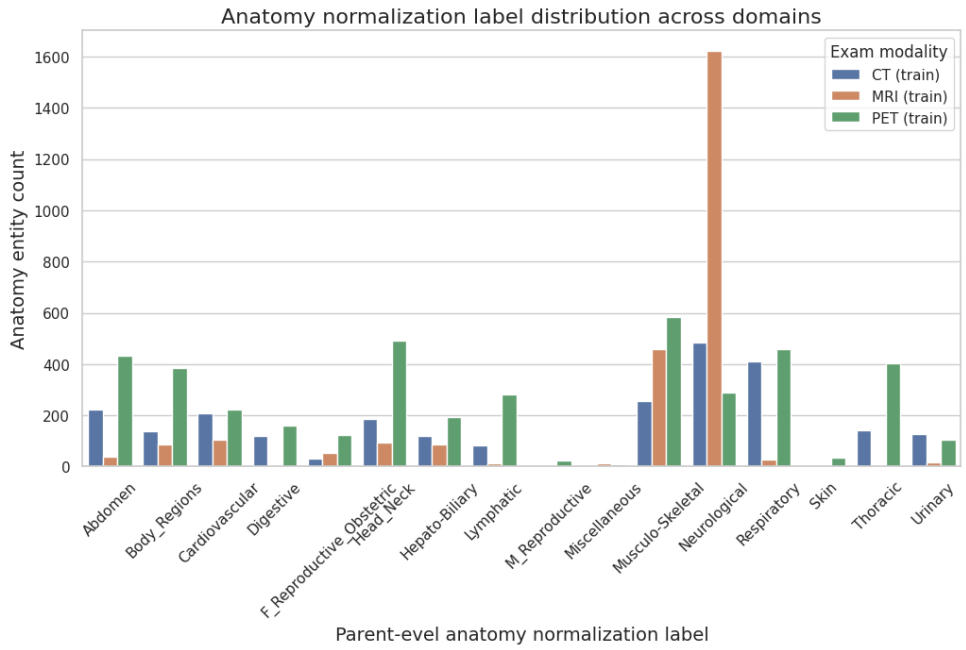

Figure 2: Domain differences in terms of the frequencies of parent-level anatomy normalization labels from the training data.

C Post-processing for the generative event extraction

When matching spans in the input sentence for predicted terms, for single-token terms, we match the corresponding token. For multiple-token phrases, we match phrases using the longest common normalized string to the input sentence. Where multiple matches are found, we choose the first match from the left of the sentence, while for anatomy spans, we choose the closest match to their query triggers.

D Domain-level context retrieval

We conduct a domain-level context search using 50,000 sentences from the target domain (PET) corpus with more than three tokens, plus 1841 sentences from the test set. The retrieved content must not be the input sentence itself. For each input clinical sentence, we identify the most lexically similar sentence from the search pool by selecting the one with the highest BM25 score. We remove punctuation and lowercase each input query when matching it with the search corpus sentences using the BM25 method.

To filter for anatomy-informative sentences, we employ the same BM25 model to match the entire search corpus with a single anatomy string, which was cheaply curated from the anatomy normalization categories and frequently auto-extracted section headers, as shown in Table 10. After filtering, the search corpus is reduced to 36%, shrinking from 51,481 sentences to 18,959 sentences.

E Implementation details

The mSpERT models are trained at a batch size of 15 for 15 epochs.[5] T5 models utilize a maximum input length of 768 tokens and a maximum output length of 512 tokens. When incorporating all types of contexts, we double the input maximum length to 1536 tokens. We train 15 epochs, with a batch size of 8. For the T5 large model, to accommodate a single NVIDIA A100 device, we employ gradient accumulation by using a batch size of 2 and accumulating four times.

F Case study for context benefits

We observe that contexts can aid in disambiguation (e.g. right middle lob) and understanding difficult medical terminology (e.g. biapical). For both examples presented in Table 11, contexts include the term "pulmonary", indicating the anatomies are related to lungs.

[5]We use full event schema for mSpERT models, including all attribute types in the annotations, including anatomy, characteristic, size, size-trend, and count. While T5 models only extract the most important attribute, the anatomy attribute.

Neurological: Spine Cervical, Spine Thoracic, Spine Lumbar, Spine Sacral, Spine Cord, Spine, Brain, Nerve, Pituitary, Cerebrospinal, Cerebrovascular, Extraaxial
Cardiovascular: Venous, Arterial, Pulmonary Artery, Heart, Pericardial Sac, Coronary Artery
Thoracic: Mediastinal
Respiratory: Lung, Pleural Membrane, Tracheobronchial
Digestive: Esophagus, Stomach, Intestine, Intestine, Intestine
Hepato-Biliary: Gallbladder, Bile, Pancreas, Liver
Urinary: Kidney, Urinary Bladder, Ureter
Reproductive: Breast, Ovary, Uterus, Adnexal, Extra-embryonic, Placenta, Fetus, Umbilical Cord, Genital Structure, Prostate, Testis, Epididymis
Musculo-Skeletal: Skeletal, Smooth Muscle, Bone, Pelvis, Limb
Head Neck: Thyroid, Neck, Ear, Eye, Mouth, Nasal Sinus, Pharynx, Laryngeal
Skin: Skin, Mucous Membrane, Subcutaneous
Abdomen: Retroperitoneal, Abdominal, Peritoneal Sac, Spleen, Adrenal, Mesentery, Adipose, Chest, Mediastinum, Osseous, Bones, Extremities, Lungs, Musculoskeletal, Ventricular, Bowel, Pleura, Spleen, Vasculature, Thorax, Gallbladder, Kidneys, Adrenals, Adrenal, Cardio

Table 10: Common anatomy terms for filtering the search scope of domain-level context retrieval. This list is curated from the anatomy task ontology (Table 8) and frequent section headers. Stop words are removed.

Table 11: Error examples with helpful contexts

Error with example	Contexts	Before and after
[ambiguity] **Right middle lobe** nodule (4, 81) measures 3 mm, previously 4 mm	[document-level section header] Scattered bilateral **pulmonary** nodules, as described below	before: Hepato-Biliary | Liver after: Respiratory | Lung
[hard vocabulary] There is **biapical** fibrosis	[domain-level BM25] There is biapical **pulmonary** fibrosis compatible with radiation therapy	before: Musculo-Skeletal | Bone and or Joint after: Respiratory Lung