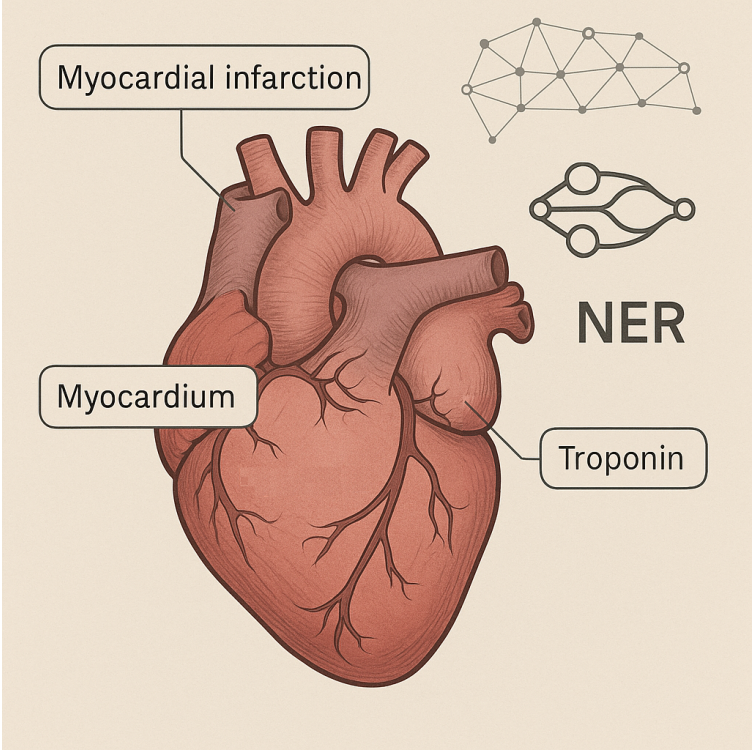


Enhancing Patient Understanding of Medical Findings Through NLP and 3D Models

Master's thesis in Data Science and AI

LUKAS ALBUSZIES

Department of Computer Science and Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
UNIVERSITY OF GOTHENBURG
Gothenburg, Sweden 2025

Master's Thesis 2025

Enhancing Patient Understanding of Medical Findings Through NLP and 3D Models

LUKAS ALBUSZIES

UNIVERSITY OF
GOTHENBURG

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Computer Science and Engineering
Chalmers University of Technology
University of Gothenburg
Gothenburg, Sweden 2025

Enhancing Patient Understanding of Medical Findings Through NLP and 3D Models

LUKAS ALBUSZIES

Supervisor: Simon Dobnik, Department of Philosophy, Linguistics and Theory of Science
Advisor: Gernot Reishofer, Medizinische Universität Graz
Examiner: Kivanç Tatar, Computer Science and Engineering

Master's Thesis 2025
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Typeset in LaTeX
Gothenburg, Sweden 2025

Enhancing Patient Understanding of Medical Findings Through NLP and 3D Models

LUKAS ALBUSZIES
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg

Abstract

This thesis investigates patient communication in clinical settings, recognizing that effective communication is vital for ensuring patient compliance and successful treatment outcomes. Complex clinical jargon and inadequately explained pathological processes often lead to patient confusion and anxiety. Advances have been made in medical text simplification, yet personalized visualizations remain underexplored. Specifically, anatomical models can provide patients with foundational knowledge about the human body, enabling clearer explanations of their medical conditions. The challenge lies in the implicit nature of anatomical references in clinical texts, where discussion focuses mainly on pathology. Anatomical entities are connected to pathological processes, but often not mentioned explicitly, for example a report on a heart attack may mention "myocardial infarction", without mentioning the associated "heart muscle". This thesis compares existing NER methods in their capability to extract implicit entities, and introduces a novel pipeline that leverages foundational biomedical entity relationships from the Unified Medical Language System (UMLS), a compendium of controlled vocabularies that provides structured mappings of relationships between biomedical entities. Extracted entities are visualized through a user interface for a 3D anatomical model, guided by a custom multi-parameter algorithm that optimizes context and clarity. Parameters include contextual distance to surrounding structures, as well as techniques for dimming, highlighting and adjusting opacity. These visualization parameters proved effective in enhancing visual representations by emphasizing relevant structures and minimizing visual clutter. An analysis on the inclusion radius of surrounding structures revealed diminishing returns for all organs tested. A custom camera positioning algorithm was used to automatically center and orient the viewpoint based on the anatomical target's bounds; this approach effectively improved the clarity and framing of visualizations. To assess annotation quality, large language models were employed as automated evaluators, scoring outputs on a five-point scale. A dedicated validation experiment demonstrated that these models could reliably distinguish between expert-curated and nonsensical annotations, supporting their use as scalable, reproducible evaluation tools. Results show that the proposed method outperforms baselines in annotation quality, with statistically significant and practically meaningful improvements. While opportunities for refinement remain, this research lays the foundation for broader applications in scenarios requiring the extraction and visualization of implicit biomedical entities.

Keywords: Data science, NER, NLP, UMLS, 3D models, Anatomy.

Acknowledgements

I am grateful to my supervisor Simon for his continuous support and insights, my examiner Kivanç for his constructive feedback, as well as my advisors Gernot and Michael for their help and their explanations for the tools explored in this work.

Many thanks also to Paul for his insights on LLMs, as well as Rodrigo and my parents Anne and Gerd for always being there throughout my endeavors.

3D renderings were created using a model by SciePro, licensed to the Medical University of Graz. The author generated all images shown in this work.

Lukas Albuszies, Gothenburg, 2025-08-14

Contents

List of Figures

List of Tables

1

Introduction

Effective communication between healthcare professionals and patients is essential for successful treatment outcomes, yet medical terminology frequently becomes a significant barrier due to its complexity and specificity [1]. Although medical dictionaries and anatomical models are widely available, they often fail to effectively explain specific diagnoses in a way laypersons can easily understand during patient consultations [2]. This communication gap significantly impacts patient adherence to treatment plans, ultimately affecting clinical outcomes [3]. Bridging this gap involves two key components: simplifying medical terms into plain-language explanations that are tailored to varying levels of comprehension, as well as intuitive visual representations such as personalized 3D anatomical models. While simplification methods are available, visualization methods remain underexplored. This thesis proposes a novel pipeline capable of extracting relevant anatomical concepts from clinical texts and dynamically visualizing them through personalized 3D anatomical models. Furthermore, factors that maximize visualization effectiveness are explored, which adds nuance to a novel approach.

Previous research in the field of medical communication highlights the challenges involved in bridging the gap between complex medical terminology and patient understanding [1]. Natural Language Processing (NLP) has been increasingly used to simplify medical jargon into layperson-friendly language, as seen in projects like Health Literacy Advisor, to emphasize readability enhancement through algorithmic simplifications [4]. Concurrently, visualization efforts, such as 3D anatomical models, have gained traction in medical education and patient consultations, with platforms like Visible Body offering detailed interactive anatomy [5]. However, these systems primarily offer generalized visualizations rather than personalized representations tailored to individual patient contexts. Addressing this specific limitation is the primary focus of this thesis.

The personalized 3D anatomical visualization of patient reports, as explored in this thesis, represents a novel and significant advancement in medical communication. While recent research has demonstrated the effectiveness of transformer-based models like BART and MUSS for simplifying complex medical text, these approaches focus solely on textual outputs at the sentence or paragraph level [6][7]. Existing models, such as Devaraj et al.'s BART-based system for medical text simplification, improve readability by penalizing jargon but do not integrate visual aids or address multimodal communication [8]. This thesis advances beyond text simplifi-

cation by integrating personalized 3D anatomical visualizations directly linked to patient reports. This approach uniquely bridges the gap between textual and visual modalities, enabling users to better understand medical concepts through intuitive visualizations [9]. Furthermore, this thesis provides an in-depth analysis of visualization techniques specifically optimized for medical communication, contributing to an area currently underexplored in existing literature. By addressing these critical gaps, this thesis addresses key limitations of prior work, offering a tool that has the potential to enhance both patient education and medical training. Integrating optimized 3D anatomical visuals represents a notable step forward in improving accessibility, clarity, and overall comprehension in medical communication.

1.1 Aim and Research Questions

This thesis aims to develop and evaluate a system that maps medical reports to a 3D anatomical model by SciePro. By addressing the complexity of medical reports, the accurate extraction of implicit anatomical entities, and the optimization of anatomical visualization, this research seeks to enhance communication within healthcare settings, thereby improving patient compliance and treatment outcomes [3]. A key component of fully realizing the potential of this pipeline, is a thorough analysis of factors that affect the clarity and relevance of visual representations. The other aspect is developing a robust pipeline that is capable of extracting implicit anatomical entities, as surface mentions mainly focus on pathology. Specifically, this research addresses three main questions:

- Does inferring anatomical sites from associated pathologies using UMLS relations improve the quality of clinically relevant anatomical site extraction from radiology reports compared to surface-level anatomical extraction, LLM-based prompting, and transformer-based entity extraction?
- Can LLM-based evaluation reliably distinguish high-quality annotations from nonsensical ones, and thus serve as a robust proxy for human evaluation?
- Does increasing the spatial inclusion radius of anatomical context lead to diminishing returns in visibility, as measured by contrast-based visibility metrics?

The central challenge addressed in this thesis is capturing and leveraging anatomical context. Medical reports frequently imply anatomical entities indirectly through mentions of pathological findings, without explicitly referencing anatomical structures [10]. The latin phrase *Anatomia clavis et clavus medicinae*, translates to 'anatomy is the key to medicine', which aligns with the structure of medical university studies, where anatomy is taught before pathology. It is unrealistic to expect laypersons to fully understand pathological findings without providing foundational anatomical context. This issue proposes the necessity of developing a system that leverages the context contained in pathological mentions, in order to link these mentions to anatomical entities as a basis for visualization. An example of a textbook visual representation of a critical condition can be seen in Figure 1.1, where the pathological process involved in a myocardial infarction (heart attack) is portrayed.

A verbal explanation of the pathological consequences of an infarction of coronary arteries may be significantly more effective in educating a patient on their condition, when combined with a visual representation of the process. This reflects the foundational role of anatomy in understanding pathological processes, which this thesis aims to leverage through contextual visualization.

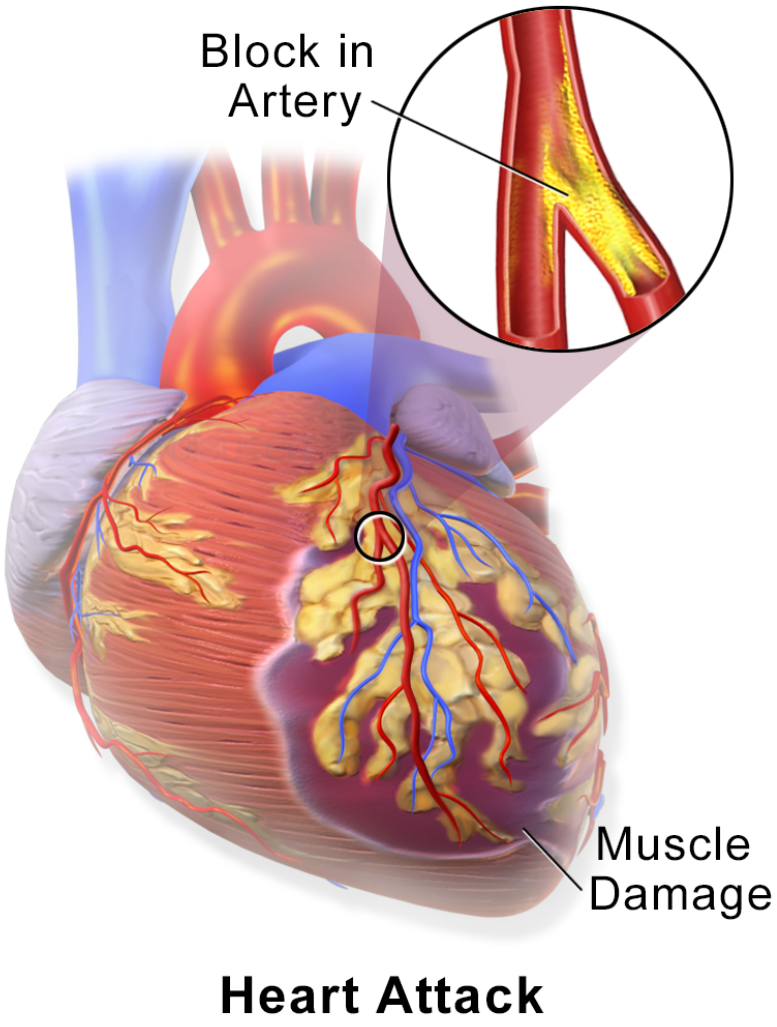


Figure 1.1: Anatomical Representation of a Myocardial Infarction. Image by Blausen Medical Communications, Inc., licensed under CC BY 3.0 via Wikimedia Commons.

1.2 Approach and Related Works

Extracting relevant anatomical entities from clinical reports fall under the category of Named Entity Recognition (NER), which is a sub task of Information Extraction (IE) within the field of Natural Language Processing (NLP) [11]. The primary goal of NER is to locate and classify named entities in text into predefined categories such as person names, organizations, locations, dates, and other domain-specific entities.

NER systems typically rely on rule-based approaches, machine learning techniques, or a combination of both. More recently, deep learning and transformer-based models (e.g., BERT) have significantly improved the accuracy of general NER tasks [12]. NER plays a crucial role in various NLP applications including question answering, knowledge base population, and text summarization by enabling structured understanding of unstructured text [11]. Recent advancements have introduced domain-specific transformer models (e.g., BioBERT for biomedical texts) and large language models (LLMs) such as LLaMA, employed in zero-shot prompting scenarios [12]. This thesis evaluates three prominent NER approaches for within clinical contexts: (1) SciSpaCy [13], which combines rule-based components with statistical modeling, (2) the transformer-based SapBERT (a domain-specific BERT) [14], and (3) LLM-based approaches using LLaMA [15] via zero-shot prompting. Clinical NER presents unique challenges within the context of the questions this thesis aims to answer, due to the complexity of medical terminology and context, implicit anatomical references within pathological references, and the limited data availability [16]. To overcome these challenges and contribute novel insights, a novel approach for extracting anatomical entities is developed in this thesis. A combination of the aforementioned tools, as well as relations between pathology and anatomy stored in the National Institute of Health's (NIH) powerful ontology for biomedical terminology called Unified Medical Language System (UMLS) [17], are employed to create a pipeline for anatomical entity extraction. The proposed pipeline integrates existing NER techniques with structured relational information from a local UMLS-based database to effectively extract implicit anatomical entities from explicit pathological mentions, and can easily be adapted to any use case that involves extracting implicit biomedical information from explicit mentions. This approach emphasizes leveraging clinical context to enhance extraction accuracy and adaptability across diverse biomedical applications [17]. The key characteristics discussed are performance and the ability to take advantage of the clinical context. The comparison of current NER techniques and evaluation of the novel approach in this thesis draws on testing with real clinical reports and large language models (LLMs) employed as automated evaluators. These LLMs were prompted to rate annotation outputs based on relevance using a five-point ordinal scale. This approach allows scalable, reproducible evaluation at minimal cost, and is validated in this work against known gold-standard and nonsensical annotations. Visualization parameters are manually assessed, and a contrast based analysis is performed on the inclusion radius of surrounding structures, to analyze whether this metric experiences diminishing returns, thereby proposing that an optimal parameter can be selected. This approach is validated through comparison to manual selection.

Related works relevant to this thesis include a review of clinical Named Entity Recognition (NER) and relation extraction techniques applied to medical free text [18], as well as specific research by Xu et al. exploring implicit anatomical references within clinical reports [10]. Xu et al. identified that numerous expressions annotated as anatomical entities in clinical texts do not explicitly refer to anatomical locations, but instead implicitly reference anatomy through associated diseases, tests, or treatments. Their research employed a hierarchical NER framework leveraging Conditional Random Fields (CRF) models and external resources such as WordNet

and Wikipedia to enhance the recognition of both explicit and implicit anatomical entities, achieving notable performance improvements (up to 5.08% increase in F1 score). However, their methodology relied significantly on external general-purpose resources such as Wikipedia rather than extensively exploiting structured anatomical-pathological relationships contained within UMLS. Building upon this identified gap, this thesis specifically utilizes the extensive, structured relationships between anatomy and pathology from the UMLS ontology, presenting a more focused approach that directly addresses the implicit extraction challenge from pathological mentions in clinical reports.

1.3 Contributions

This thesis makes the following key contributions:

1. A novel pathology-driven pipeline for extracting clinically relevant anatomical sites from radiology reports, which infers anatomical targets by linking identified pathologies to anatomical structures using SNOMED CT relations.
2. A comprehensive empirical comparison of this pipeline against three alternative approaches: surface-level entity extraction, LLM-based prompting, and transformer-based anatomical extraction.
3. An evaluation framework using LLMs as automated judges, validated through a discrimination task between gold-standard and nonsensical annotations, demonstrating that large language models can reliably assess annotation quality at scale.
4. An integration into a Unity-based clinical interface with several visualization parameters, illustrating the feasibility of embedding semantic extraction pipelines into clinical tools.
5. An analysis on the diminishing returns of added surrounding structures measured by contrast based metrics.
6. An analysis on the reliability of contrast based metrics to serve as a proxy for human visibility.

1.4 Summary

This thesis addresses the communication gap between complex medical terminology and patient comprehension by integrating clinical reports with personalized 3D anatomical models. A novel framework is developed that leverages advanced Named Entity Recognition (NER) techniques to extract implicit anatomical information from medical texts, linking these directly to intuitive 3D visualizations. By analyzing visualization perspectives, the thesis tackles significant medical communication challenges, including diminishing returns of added context and clarity of visual representation. The proposed tool aims to enhance patient understanding and improve clinical outcomes, offering broad applicability in healthcare communication

and medical education. Future work could include multilingual adaptation and integration with existing medical text simplification techniques, further extending the potential impact of this research.

2

Theory

This chapter explores the current tools and methods available for improving patient understanding, and highlights the importance with reference to improved treatment outcomes. Specifically, it provides a theoretical comparison of current Named Entity Recognition (NER) techniques used in clinical Natural Language Processing (NLP), evaluating their strengths and limitations in the context of anatomical entity extraction, as well as a comprehensive review of advancements and results in improved patient communication. Additionally, it introduces the Unified Medical Language System (UMLS), a foundational biomedical ontology used in this thesis to link pathological mentions to anatomical structures. Furthermore it provides context on the way contrast and luminance interact with human perception in images.

2.1 Comparative Analysis of Clinical NER Methods

2.1.1 Introduction

A core focus of this thesis is Clinical Named Entity Recognition (NER), the task of identifying and extracting medical concepts such as diseases, treatments, and medications from clinical narratives. This task presents unique challenges due to the complexity of medical terminology, the nuanced clinical context, and the limited availability of annotated data. Recent advances in clinical NER span a spectrum of methodologies, including rule-based systems, transformer models, and Large Language Models (LLMs). This thesis evaluates and compares three approaches for clinical NER: (1) the SciSpaCy pipeline (which combines rule-based components with statistical modeling), (2) transformer-based models like BioBERT (a domain-specific BERT), and (3) LLMs such as LLaMA, used via zero-shot prompting. The evaluation centers on two primary criteria: NER performance and the ability to utilize clinical context effectively. The comparison draws on recent benchmarks and studies (2023—2025), and a summary of key metrics and integration considerations are included in Table 2.1.

2.1.2 SciSpaCy Pipeline NER

SciSpaCy is a spaCy-based NLP toolkit tailored for biomedical text [13]. Its pipeline includes a pre-trained spaCy model and optional rule-based components. For exam-

ple, SciSpaCy can detect common abbreviations via rule-based patterns and link entities to a UMLS dictionary. The statistical NER models in SciSpaCy are trained on biomedical corpora (e.g. JNLPBA, BC5CDR), achieving respectable accuracy (e.g. F1 72—85% on biomedical entity benchmarks [19]). For instance, on the BC5CDR corpus (chemical and disease names), a SciSpaCy NER model achieves about 85.5% F1. This is within a few points of transformer models on the same data (BioBERT reaches 87.8% F1 on BC5CDR) [19], showing that SciSpaCys pre-trained models are competitive for basic biomedical entity extraction. However, contextual understanding in SciSpaCy is more limited. The spaCy NER uses a convolutional or transition-based model (or a transformer if using SciSpaCys scibert model) with a context window covering a single sentence. It may not capture long-range dependencies as effectively as transformer encoders. For example, SciSpaCys performance drops on tasks with many entity types and complex contexts: in one benchmark with a wide variety of biomedical entities, SciSpaCys F1 was 78% versus BioBERTs 86% [19]. SciSpaCys built-in abbreviation resolver and ontology linker do help with domain-specific vocabulary (e.g. mapping HTN to hypertension), but they rely on exact or fuzzy string matches rather than true semantic disambiguation. Thus, SciSpaCy might mislabel an ambiguous term if the context is not strong, whereas a transformer model could use surrounding context to decide if, for example, Mass refers to a tumor or a measurement unit. On the other hand, SciSpaCys reliance on curated ontologies is a strength for precision: it tends to detect entities that match known medical terms, reducing spurious outputs. But this can come at the cost of recall if an entity is not in the lexicon or the statistical model has not seen a similar term. Fine-tuning SciSpaCy on a specific clinical dataset can improve its recognition abilities [2], but this requires additional annotation and training with spaCy. SciSpaCys major advantage is efficiency and ease of integration. It is lightweight and can run on CPU, making it attractive for deployment in hospital settings where GPU resources might be limited. Its inference speed is significantly faster than large transformers: one study found SciSpaCy processed a sentence in 90 ms on average, versus 278 ms for BioBERT [19].

This difference highlights the trade-off between speed and accuracy. Integrating SciSpaCy is straightforward as it is available as a Python package with pre-trained models. The weaknesses of SciSpaCy lie in its slightly lower accuracy in complex cases and its need for additional configuration to cover new entity categories (since its models are trained on fixed sets of entity types), as well as reliance on surface level mentions. In summary, SciSpaCy is a fast, ready-to-use pipeline that performs well on many biomedical NER tasks, though it may not perform as well as context-aware models on clinical text.

2.1.3 Transformer-Based Models (BioBERT and Variants)

Transformer models pre-trained on biomedical text have become the current standard for high-performance NER in clinical domains. BioBERT for example, is a BERT-base model further pre-trained on large biomedical corpora (PubMed abstracts and PMC articles) [20]. This domain-specific pre-training gives BioBERT an understanding of medical vocabulary and context. BioBERT achieves higher

NER accuracy than general BERT on all tested biomedical datasets, it can recognize biomedical entities that a general BERT model often misses [21]. This indicates improved contextual understanding of domain specific terms. Fine-tuning is important to achieve the highest possible NER performance. A pre-trained model like BioBERT or ClinicalBERT (BioBERT further adapted to clinical notes) must be fine-tuned on an annotated NER dataset (such as i2b2 clinical notes or NCBI disease corpus) to learn the specific entity categories and span boundaries. Fine-tuned BioBERT models have achieved impressive results on many benchmarks. For example, a BioBERT model fine-tuned on the NCBI Disease corpus reaches F1 of 0.89 [22]. In the clinical domain, BioClinicalBERT (a variant initialized from BioBERT and trained on MIMIC-III clinical notes) obtained F1 0.90 on the 2010 i2b2 clinical concept extraction task [23]. These high scores (approaching 90—91% F1) significantly outperform earlier rule-based or statistical approaches, demonstrating excellent precision and recall. Even on more complex NER tasks (e.g. recognizing medication names and related attributes from discharge summaries), fine-tuned BioBERT models have reported F1 around 0.91 [19], showcasing their understanding of clinical context. The pre-trained BioBERT without fine-tuning is not directly usable for NER, however, the benefit of its pre-training is seen in the fine-tuned results, which are typically a several point F1 improvement over models pre-trained on only general text [21].

Transformer NER models like BioBERT excel at leveraging context within their input window (usually up to 512 tokens). They consider the full sentence (or multiple sentences if within the token limit), enabling them to resolve ambiguities using context. For example, if a note says Pt with cold was prescribed cold compress, a fine-tuned BioBERT can use surrounding words to decide that the first cold is a problem (illness) while the second refers to temperature (not an entity). This is a key advantage over simple dictionary matching. Long-range dependencies beyond the 512-token limit are therefore a weekness in this method. If a clinical note is very long, it may need to be split into segments for the transformer model, which could lead to missing some cross-segment context. Nonetheless, within a given window BioBERT captures more context than traditional models, due to its self-attention mechanism. Domain-specific transformers are also robust to synonyms and variants. For example, BioBERT has likely seen both myocardial infarction and heart attack in pre-training, so it can identify either as a cardiac event entity once fine-tuned, even if one phrasing was scarce in the fine-tuning data. Another weakness is the requirement of labeled data for each new NER task. Training data is needed to fine-tune, which is scarce in the clinical domain and can be costly to obtain. Moreover, transformers are heavier than rule-based pipelines and while not as resource-intensive as full LLMs, BioBERT has approximately 110M parameters, likely needing a GPU for efficient inference. Despite this, integration is fairly straightforward and many pre-trained weights (e.g. BioBERT, ClinicalBERT) are openly available, and fine-tuned versions for common datasets can be downloaded [24]. Deployment in clinical settings is feasible on modern hardware, and latency is acceptable for most applications with only hundreds of milliseconds per sentence [19]. In summary, transformer-based models (especially domain-specific ones like BioBERT) offer a balance of contextual understanding and performance for clinical NER, at the cost of needing task-specific

fine-tuning and moderate computational resources.

2.1.4 Large Language Models (LLMS: e.g. LLaMA) for NER

Large Language Models (LLMs) such as OpenAI's GPT-4, and Metas LLaMA represent a new paradigm in the field of NLP: these models are pretrained on massive corpora (including biomedical text) and can perform tasks via prompting, without explicit fine-tuning for each task. Their potential for zero-shot or few-shot NER is of great interest for this thesis given the limited availability of annotated clinical notes. However, current evidence shows that out-of-the-box LLMs still trail fine-tuned biomedical models in NER performance [24] [23]. For example, on a clinical note concept extraction task, GPT-4 achieved an F1 around 0.80 with basic prompting, and about 0.86 F1 with extensive prompt engineering and few-shot examples. This was an improvement over GPT-3.5 (which reached 0.79 F1 with the same prompt techniques) [23], and demonstrates how careful prompting can enhance LLM accuracy. However, even GPT-4s best performance (86.1% F1) was lower than that of a dedicated BioClinicalBERT fine-tuned on that task (90.1% F1). In another benchmark (NCBI Disease NER), GPT-4 in a zero-shot setting reached only 59.9% F1, roughly 30% (absolute) lower than the state-of-the-art fine-tuned approach (which exceeded 90% F1) [22]. These results show that without task-specific tuning, LLMs may miss many entities. In short, a large models sheer knowledge does not automatically translate into perfect entity extraction; they benefit from guidance (prompts or examples) to focus on the task. One way to close this gap is fine-tuning LLMs on NER data. Open-source models like LLaMA-2 can be fine-tuned or instruction-tuned for NER. This dramatically boosts their performance, at the cost of significant computational expense. Fine-tuning a LLaMA-2 (13B) on a NER dataset yielded F1 0.868 on NCBI Disease, approaching BioBERTs level (0.892) [22]. With enough data, a fine-tuned LLaMA model can slightly surpass smaller models: one study found an instruction-tuned LLaMA-3 (70B) model outperformed a BERT-based model by 7% F1 on an unseen clinical NER dataset [25]. The trade-off, however, is huge, LLaMA required far more memory and was up to 28 times slower in inference. This difference is crucial in determining the best suited model for clinical NER. A 70B-parameter model for NER in real clinical pipelines may be impractical when a 110M-parameter BioBERT can do nearly as well. However, LLMs have several advantages that provide utility. They have strong generalization, meaning that in low-data settings, an LLM can use its pretraining knowledge to identify entities that a smaller model might miss. LLMs are also function with long-range context: with context windows of thousands of tokens, models like GPT-4 can process an entire clinical report and link information across the document when extracting entities. This is advantageous for complex clinical notes where relevant context for an entity (such as an abbreviation definition or a negation cue) might be paragraphs apart. Traditional NER models, limited by input length, could miss such connections. In terms of contextual understanding, large models have a strong grasp of nuance. This means they can not only understand medical terminology but can follow task instructions in prompts and adapt its out-

put accordingly [23]. Ambiguity resolution also works very well, as an LLM might use its general world knowledge to infer what is meant with ambiguous terms. Aside from aforementioned computational costs, LLMs have other weaknesses. They can hallucinate text that wasnt in the input, especially if the prompt is not well-crafted [22]. Furthermore, generative models struggle with consistent boundaries. If not constrained, they might merge two adjacent entities or split one entity into parts inconsistently. Prompt engineering or using an extraction paradigm (like instructing the LLM to output a JSON of entities with character offsets) can improve on this. Additionally, closed-source LLMs raise data privacy and integration concerns in healthcare, as sending sensitive clinical text to an external API may be forbidden. Open-source LLMs (LLaMA, etc.) avoid this but require heavy infrastructure to deploy locally. A 13B model can sometimes be run on a single GPU, but anything larger (30B, 70B) might need multi-GPU servers, which is a significant deployment investment. Maintenance and updates are also non-trivial, whereas a smaller model like BioBERT can be versioned and managed more easily. To summarize, LLMs in NER offer unparalleled flexibility and contextual reasoning, with the ability to operate in a zero-shot manner on new entity types or schemas simply by changing the prompt. They also can handle longer documents as a whole. However, in pure performance terms, a fine-tuned domain-specific BioBERT still often has the edge in F1-score for clinical NER, at a fraction of the computational cost.

The gap is closing as prompt strategies improve and as domain-specific LLMs emerge (e.g. models like PMC-LLaMA which are further pre-trained on biomedical literature [22], or instruction-tuned clinical LLMs like Clinical Camel and others). Future LLMs specifically trained on medical text may achieve high NER performance without full fine-tuning, but at present the practical choice leans toward fine-tuned moderate-sized models for efficiency. The decision between using an LLM versus a traditional model for NER should consider the use-case: if maximal accuracy on a specific dataset is needed and labeled data is available, fine-tuning BioBERT/ClinicalBERT will yield excellent results. If the task requires specific handling and instructions, a lightweight LLM may be the better choice.

Table 2.1: NER Method Comparison in Clinical Texts

Method	**Contextual Ability**	**Strengths/Weaknesses**	**Deployment**
SciSpaCy	Local sentence-level; limited deep context	+ Fast, rule-based, precise on known terms – Weak on ambiguity and novel terms	Easy (lightweight, CPU-friendly)
BioBERT	Strong local/global context; 512-token window	+ Accurate, domain-tuned, handles variation well – Needs fine-tuning, GPU preferred	Moderate (popular libraries support it)
LLMs (LLaMA)	Excellent long-range and doc-level context	+ Flexible, zero/few-shot capable – Resource-heavy, possible hallucinations	Hard (needs GPUs, careful integration)

2.1.5 Evaluation of NER Methods with LLM Judges

Large language models like Mistral and LLaMA3 can serve as automated judges to evaluate the outputs of Named Entity Recognition systems [26]. In this setup, an LLM is prompted with the original text and the NER systems output (e.g. a list of extracted entities with their types) and tasked with assessing the outputs correctness and completeness. For example, a LLaMA3-70B or Mistral model can be instructed to verify whether each identified entity actually appears in the text with the right label, and whether any entities were missed or mislabeled [26]. The LLM judge may produce a score or verdict (often on a numeric scale or as a pass/fail judgment), sometimes accompanied by a brief explanation. Two prompting strategies are common: pointwise evaluation (scoring a single NER output against criteria like accuracy and recall of entities) and pairwise comparison (given two NER outputs for the same text, deciding which one is more correct) [27]. Carefully crafted prompts are used to guide the model to check the text and output before making a judgment, which helps improve its evaluation reliability [28]. Using LLMs as NER judges offers a scalable alternative to human evaluation, and studies show that well-designed judge models can align with human judgment [27]. In fact, the largest models (e.g. LLaMA3 70B) tend to achieve the highest agreement with human evaluators on similar tasks [27]. They can leverage their broad knowledge to handle tricky cases, for instance, recognizing a nickname or abbreviation as an entity that a simpler automatic metric might miss. However, there are important limitations and potential pitfalls as LLM judges can exhibit biases. For example, they might favor an output that looks more thorough (longer lists of entities) or be influenced by the order in which outputs are presented in a comparison [26]. They may also miss subtle context or nuances that a human would catch, or even hallucinate errors, such as mistakenly insisting an entity is incorrect or missing based on the models own world knowledge rather than the given text [27]. This echo chamber effect (favoring responses that resemble the models training data or expectations) is a known risk when using AI for evaluation [29]. Additionally, smaller models or insufficiently tuned judges often show lower alignment with human ratings, and even the best LLM judges still fall short of perfect agreement with humans [27]. Handling edge cases (such as very long documents or ambiguous entity mentions) also proposes a challenge, as LLM judges might become inconsistent or overly lenient in such cases if the prompts are complex or the context is extensive [27]. For these reasons, it is crucial to design the evaluation prompts and criteria carefully and to validate the LLMs judgments against human reviews [30]. When properly aligned with human-defined criteria, LLM-based evaluation can efficiently flag NER errors and approximate human judgment.

2.2 Unified Medical Language System (UMLS)

The Unified Medical Language System (UMLS) is a foundational resource developed by the U.S. National Library of Medicine (NLM) with the aim of enabling interoperability among diverse biomedical and clinical information systems [17]. It serves as an integrative framework that supports the mapping, translation, and semantic

alignment of a wide variety of medical terminologies, thus playing a critical role in biomedical informatics and health information technology. UMLS was designed to address the challenge of semantic heterogeneity in healthcare and biomedical data. It provides a standardized means of linking different vocabularies, terminologies, and coding systems, facilitating data exchange and consistent interpretation across platforms and institutions. This capability is particularly essential in contexts such as electronic health records (EHRs), clinical decision support systems (CDSS), and medical research databases. An overview of knowledge sources included in UMLS, such as SNOMED CT and MeSH, is shown in figure 2.1. The sheer amount of information stored in this system is a result of the core concept behind UMLS: the combination of several biomedical knowledge sources such as SNOMED CT. The entities extracted from these sources, are outlined in figure 2.2. At the heart of UMLS is its Semantic Network, which organizes biomedical entities and their relationships. These relationships include semantic links such as *partof*, as illustrated in Figure 2.3. For example, a "Cell" may be connected to a "Tissue" through such a relation. This relational structure is central to this thesis, as it enables the linking of pathological contexts in clinical texts to the implied anatomical entities.

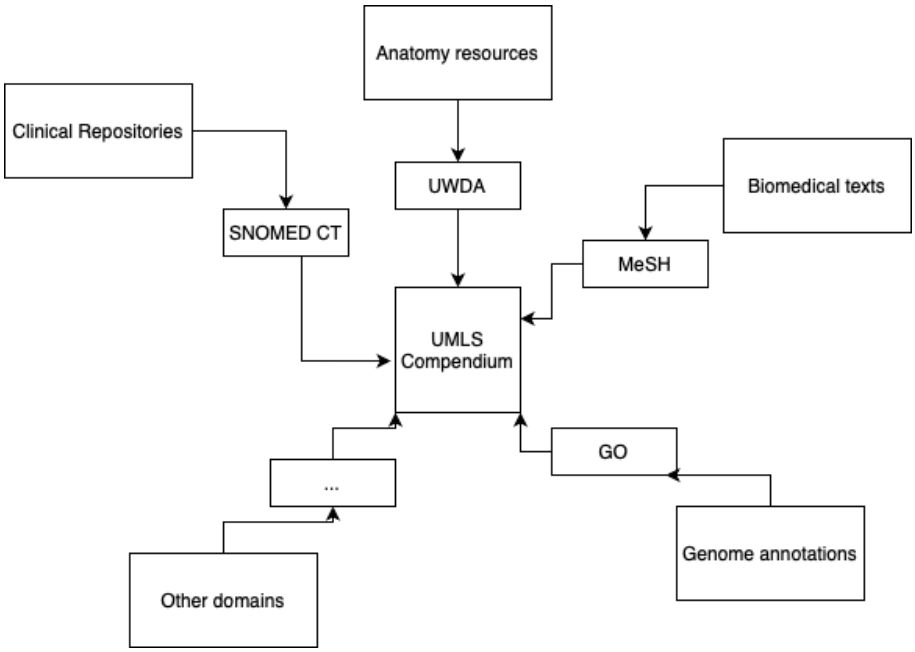


Figure 2.1: Overview of UMLS Knowledge Sources

2.3 Current Methods for Improving Patient Understanding

The current advances in improving patient communication with the goal of improving understanding and therefore treatment outcomes focus mainly on medical text simplification. Because of the lack of research in the field of personalized visualisations, a brief exploration of the benefits of improvements in patient outcomes with medical text simplification is outlined here.

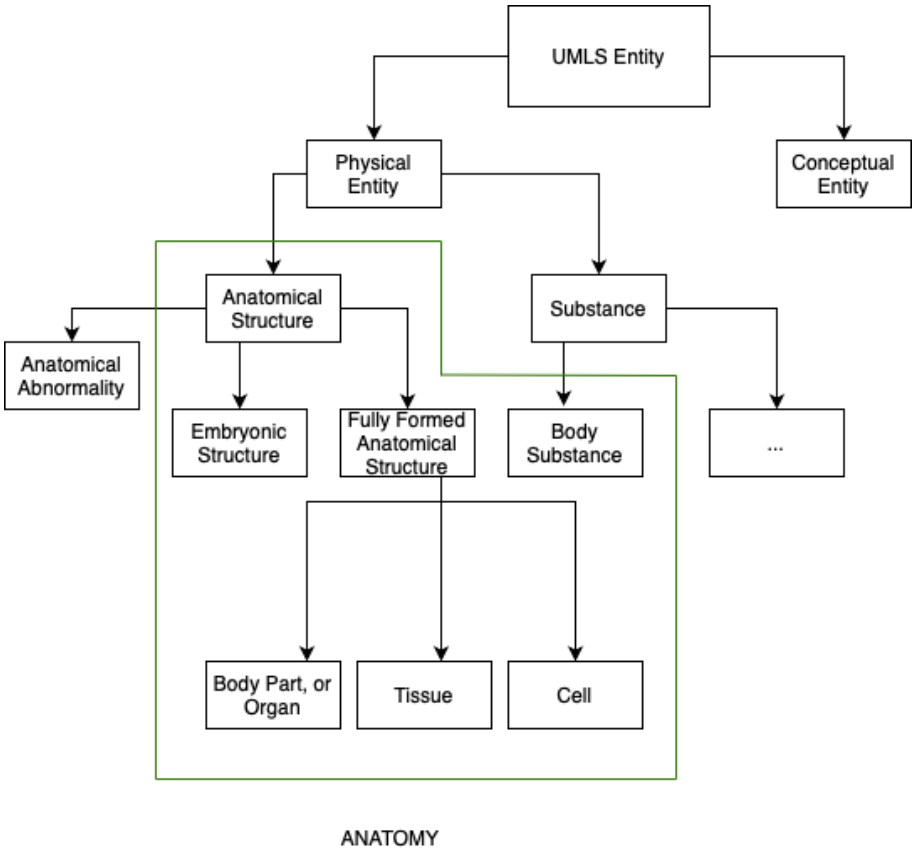


Figure 2.2: Hierarchal Representation of Entity Types

Medical text simplification aims to rewrite complex clinical content such as test results, discharge summaries, and doctors notes into plain language that patients with little medical background can understand. This is important as these documents are often filled with jargon, abbreviations, and technical detail. In fact, most clinical reports are written for other healthcare professionals; one study found only about one-third of hospital discharge summaries contained any patient-centered information [32]. The result is widespread misunderstanding. For example, one study found that 76% of parents of children undergoing an endoscopy did not understand alternatives to the procedure [33], where the cause was found to be inadequate explanation of medical results [34]. Simplifying medical text addresses this gap by improving comprehension, which in turn can improve patient outcomes. Research shows that using patient-friendly language is associated with lower readmission rates and fewer follow-up calls from confused patients [35]. Health literacy guidelines now explicitly recommend reducing medical jargon and using everyday language to communicate health information [32]. This concept aligns with the task explored in this thesis. The current state of the art in medical text simplification will be reviewed, focusing on NLP-based technologies, and how they cater to different patient groups, how they are evaluated for comprehensibility, and their impact on patient outcomes. Key developments from the last five years will be reviewed and an examination of which solutions are proving practical in clinical workflows (e.g. patient portals, EHR-integrated tools, mobile apps) will be performed.

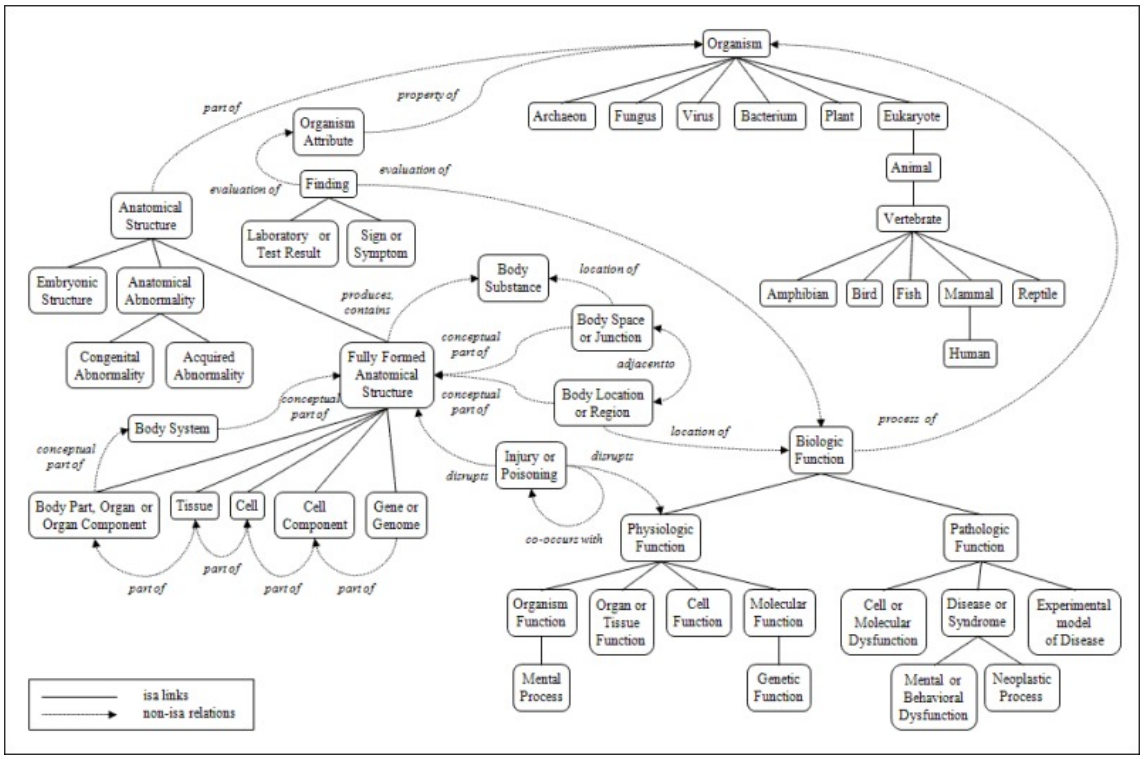


Figure 2.3: Portion of the UMLS Semantic Network Relations. Source: U.S. National Library of Medicine, Unified Medical Language System (UMLS) [31]

2.3.1 Technologies and Techniques for Simplification in Clinical Settings

2.3.1.1 Lexical Simplification

A fundamental step in medical text simplification is substituting complex terms with simpler ones. An existing tool that links medical jargon to lay definitions is NoteAid, which has improved comprehension for health records among crowdsourced participants [36]. Tools based on terminology substitution, however, do not guarantee improvement [37]. This suggests an enhanced understanding requires addressing the broader context of the report.

Simplification systems rephrase complex sentences. For example, a sentence like "No acute cardiopulmonary abnormality is defined" may be rephrased as "We did not find any new problems with your heart or lungs". Studies have explored the impact of such changes, splitting complex noun phrases can help readers, though benefits vary [37].

Rule based systems use predefined templates or rules to present information in simpler terms. For example, a lab report might auto-generate a phrase like "Your LDL cholesterol is 160 mg/dL, which is higher than the normal range (up to 130 mg/dL)" instead of a brief "LDL: 160 (H)". Rules can also expand abbreviations (turning "BP" into "blood pressure") and remove technical detail. An early system by Kandula et al. (2010) did exactly this, simplifying health education materials

by removing complex phrases and adding explanations [38]. Rule-based approaches are deterministic and easily controllable, but they require significant expert effort to develop and may not cover every nuance of free-text narratives.

Fine-tuned models like BERT, T5, and BART have been used on medical simplification tasks. These models bring powerful language generation capabilities and can incorporate medical context if further trained on biomedical text. For example, fine-tuning a T5 model on paired expert-lay data can yield a simplification system that rewrites input sentences into a targeted reading level. A challenge specific to medicine is ensuring accuracy, a simplification model must not distort or omit critical facts (a risk known as losing meaning preservation). Evaluation campaigns (e.g., in the 2022 COLING workshop) compared algorithms on not just readability, but also on how well they preserve the original information [37]. Pure statistical (non-neural) approaches are rare in biomedical simplification, likely due to the limited data availability.

When GPT-style large language models became popular, they were applied to medical text simplification. These models (e.g. GPT-3, GPT-4) are trained on massive amounts of general text and can perform tasks like simplification via prompting, even without specialized training data. Early studies tested models like ChatGPT on real clinical text. For example, Jeblick et al. (2023) prompted ChatGPT to simplify radiology reports for patients [39]. In an assessment by radiologists, the majority of ChatGPTs simplified radiology reports were rated factually correct, complete, and not harmful to patients, though some mistakes were present. Another study by Lyu et al. (2023) used GPT-4 with specialized prompt tuning to translate radiology findings into plain language, reporting promising results with some limitations [32]. AI outputs are indeed shorter and simpler (average reading level 10th grade vs 10.7 originally, and using fewer complex words), but around 18% of outputs had potentially unsafe issues (e.g. hallucinated facts or incorrect advice). [32] So although current LLMs can produce fluent and accessible text, they require careful validation in the medical context, to avoid dangers. Prompt engineering may improve results.

2.3.2 Evaluation Metrics

Traditional readability metrics, such as Flesch Reading Ease, FleschKincaid Grade Level (FKGL), SMOG Index, Gunning Fog Index, and ColemanLiau Index, quantify the reading level of a text [37]. These formulas use features like word length, sentence length, and syllable counts to estimate how difficult a text is. In the GPT-3.5 discharge summary study, the original notes averaged roughly a 10th-grade level, whereas the AI-simplified versions averaged around 10.1 (slightly lower, indicating somewhat simpler text) [32]. However, these metrics do not account for medical jargon or context. A sentence could score as easy by grade level (short words and sentences) but still be incomprehensible if it contains unfamiliar medical terms. Researchers say that readability scores can be misleading for automatic text simplification [37], and should be used alongside other measures. In practice, many simplification studies report FKGL or SMOG for reference, but they increasingly

rely on additional metrics.

When parallel reference simple texts are available, machine translation-style metrics can be used to compare the AIs output to these references. A widely used one is SARI (System output Against References and against the Input) [37]. SARI measures how well the system adds, deletes, and keeps content relative to the original and the reference simple text. Other MT metrics like BLEU and ROUGE (which measure n-gram overlap with references) are sometimes reported, but they have drawbacks. A simplification can be good even if it uses different wording than the reference (which might lower BLEU), and BLEU/ROUGE dont directly capture simplicity or readability. In fact, if a model simply copies the original text (which is perfectly grammatical and on-topic), BLEU could be high, but the output would not be simplified at all. For that reason, SARI is preferred in recent simplification evaluations as it penalizes not simplifying enough and also penalizes over-simplifying). In the Med-EASI work (Basu et al. 2023), for instance, SARI and related content-preservation metrics were used to tune models for an optimal simplicity-vs-accuracy tradeoff.

2.3.3 Impact on Patients and Care

2.3.3.1 Improved Comprehension

Simplified texts have been shown to directly increase what patients remember and understand about their condition and care. A quality-improvement study in an emergency department created a one-page simplified discharge instruction sheet covering diagnosis, treatment given, home care, follow-up, and return precautions. They found that patients knowledge scores improved by 22% after this intervention [40]. Notably, no patients in the initial assessment had fully understood all aspects of their ED visit and instructions, underscoring how standard materials were falling short. After implementing the simplified instructions (called the SIP, Simplified Information Page), patients across all demographic groups showed better comprehension, meaning the solution was broadly effective [40].

2.3.3.2 Treatment Adherence

When patients understand instructions, they are more likely to follow them. This includes taking medications correctly, doing recommended exercises or wound care, and attending follow-up appointments. For example, one of the lowest-understood items is medication frequency and duration [40], which is critical for adherence. By simplifying instructions and information about patients' conditions, adherence can improve. One specific intervention, the Universal Medication Schedule (UMS) format, was designed to simplify medication instructions and has been associated with improved medication adherence, particularly in older patients [32]. Patients are less prone to make dosing errors when instructions are written in plain language tied to daily routines. This suggests that when instructions are clear, and comprehension of conditions is improved, patients manage better on their own, reducing complications that lead to hospital returns.

2.4 Human Perception and Contrast

In the context of 3D anatomical visualization, visual salience of a target structure depends critically on its perceptual separation from surrounding anatomy [41]. One of the most fundamental cues for visual separation is luminance contrast, which determines how easily a viewer can distinguish a target from its background [42]. This principle underpins the design of most visibility metrics in medical imaging and scientific visualization.

The human visual system is acutely sensitive to differences in brightness, with contrast functioning as a low-level visual feature that strongly guides attention [43]. Higher contrast between a target and its background generally improves detection and recognition performance, particularly under time pressure or in cluttered environments [44]. Traditional contrast models such as Weber and Michelson contrast have been widely used to quantify this effect [45]. However, these models often rely on assumptions about background luminance dominance or symmetry that may not hold in complex, multi-structure visualizations.

3

Methods

3.1 Overview

The central issue this thesis tackles, is that the anatomical entities relevant to diseases described in medical reports are often implied through discussion of pathological processes [10]. For example, a report on a heart attack would mention "myocardial infarction", as well as the enzymes measured in the patient's blood to confirm the necrosis of myocardium. It may also mention pain experienced by the patient radiating to the left arm. A rule-based NER method trained on surface level mentions as discussed in the theory chapter would label "left arm" as the anatomical entity, which does not provide useful information about the main issue the patient has. Instead, a functioning pipeline would output the affected areas of the heart, such as the myocardium, or the coronary arteries, see Figure 1.1 for reference. This proposes the need for a novel approach that leverages biomedical entity relations, tailored to the use case explored in this thesis. First, out of the box NER techniques utilizing rule-based, transformer-based and LLMs are compared, and then a functioning pipeline that employs a combination of NER techniques, as well as UMLS relations is developed. This pipeline comes with the added functionality of adaption to extracting any desired type of implied information from explicit mentions in the biomedical context. This pipeline is packaged and setup for a backend rest API, in order to call it in Unity. To determine the extent to which this method provides more meaningful results over traditional NER methods, a comparison based on LLM evaluation is performed, which is validated by gold standard and nonsense datasets. To explore the visualization techniques available, and determine their effectiveness in providing context while maintaining clarity, an algorithm for anatomical entity visualization is developed. The Unity-based visualization component involves a custom multi-parameter algorithm that includes parameters optimized for contextual distance to surrounding structures, as well as techniques for dimming, highlighting and adjusting opacity. It furthermore selects the correct camera position based on the size and position of the object. The context distance is evaluated on its diminishing returns, which is validated through comparison to manual selection. A user interface for report input and visualization triggering is implemented to provide an all-in-one system.

3.2 Dataset Preparation

The dataset used in this thesis is a subset of MIMIC IV radiology notes [46]. It consists of 1100 reports, and has been filtered for the 'Impression' section, which is a summary of the key findings. The baseline NER techniques, as well as the proposed pipeline were evaluated on the full 1100 reports with the LLMs LLaMA 3 and Mistral as judges. For the specific use case of evaluating implicit anatomical entity detection, no gold standard annotations were available. Instead a manual annotation of 100 sample MIMIC IV notes, as well as a nonsense dataset with 100 random words were used to validate the evaluation method.

3.3 Anatomy Extraction Approaches

3.3.1 Baseline Approaches

3.3.1.1 SciSpacy

To explore the capabilities of statistical Named Entity Recognition (NER) systems in extracting underlying anatomical entities, SciSpacy's en_core_sci_sm is used to extract biomedical entity mentions [13]. This model was selected for its strong baseline performance in biomedical NER tasks and ease of integration [19]. Although en_core_sci_sm is optimized for general biomedical entity recognition rather than task-specific extraction, it provides a reliable starting point for evaluating how off-the-shelf models handle anatomical references, both explicit and implicit, within pathology-focused clinical narratives. The model provides filtering for semantic types. In the experimental setup, this filter was set to the Type Uniques Identifiers (TUIs) outlined in table 3.1. These are the TUIs that correspond to anatomical entities, and therefore the model output is filtered to only return the anatomical entities that were detected.

Table 3.1: UMLS Semantic Types used for Scispacy Extraction

Code	Semantic Type Name	Category
T017	Anatomical Structure	Anatomy
T029	Body Location or Region	Anatomy
T023	Body Part, Organ, or Organ Component	Anatomy
T030	Body Space or Junction	Anatomy
T021	Fully Formed Anatomical Structure	Anatomy

For example, in the sentence "A patient presents with pneumonia", SciSpacy would detect the explicitly mentioned entities "Patient" and "Pneumonia" as biomedical entities. The filter for anatomical entities would discard these entities as they do not correspond to anatomical entities. In the sentence "A patient presents with pneumonia, following a viral infiltration of the lung", the entity "lung" would be detected along with the others and survive the filter, since this is an anatomical entity. Performance in recognizing entities such as diseases, symptoms, and anatomical terms

serves as a benchmark against which transformer-based and LLM approaches are compared later in this study.

3.3.1.2 SapBERT

To evaluate the entity recognition capabilities of transformer-based models, this thesis uses SapBERT-from-PubmedBERT-fulltext, a domain-specific variant of BERT designed for biomedical concept extraction [14]. SapBERT extends the standard BERT architecture by incorporating self-alignment pretraining, which encourages semantically similar biomedical terms, such as synonyms from UMLS, to be mapped to nearby points in the embedding space [47]. The specific variant used in this thesis is pretrained on full-text articles from PubMed and fine-tuned using the 2020 UMLS release, aligning well with the biomedical scope of this thesis. The model was implemented according to the procedure outline in Algorithm 1.

Algorithm 1 SapBERT Pseudocode for Embeddings Based Entity Matching

Require: embeddingsFile, reportText or reportFile, modelName, maxLength, topK
1: device accelerator or CPU
2: Load tokenizer & model(modelName) on device
3: Load and normalize term embeddings from embeddingsFile
4: Read report from reportText or reportFile
5: Embed report + mean-pool, then normalize
6: Compute cosine similarities
7: Get topK indices & scores
8: Print terms with scores

Each model label entry from a list of all object labels for the 3D anatomical model (e.g. "Left Kidney") is embedded using only its name, and stored in a serialized .pkl file for efficient access. To match entities from clinical text, the entire clinical report is embedded using the same model. Cosine similarity is then calculated between the report embedding and all label embeddings. The entry with the highest similarity score is then returned. This approach enables semantic matching between varied clinical phrasing and anatomical concepts.

3.3.1.3 LLaMA

LLMs, as discussed in the Theory chapter, have shown strong capability of extracting nuanced information from clinical reports. While transformer-based models are effective at identifying semantically similar terms through embedding-based proximity, they lack the capacity for nuanced reasoning or prioritization. Therefore, the LLaMA 2 large language model is introduced at this stage to evaluate the capabilities of LLM's for implicit entity detection [15]. The importance of testing a locally installed version lies in the sensitivity of medical data, due to which an online solution such as OpenAI's API is most likely not viable in a clinical setting.

In this implementation, LLaMA receives a prompt consisting of the full clinical report along with a structured instruction to return the single most relevant anatomical entity based on the clinical context. This setup enables the model to handle

reports that mention multiple anatomical structures (e.g., "aortic valve" and "left ventricle") and prioritize based on pathological relevance rather than surface-level proximity in the text or embedding space. Specifically, the prompt goes as follows: "Return the single most relevant anatomical entity in the following report: [REPORT] Reply with the anatomical entity only."

3.3.2 Novel Proposed UMLS-Enhanced Pipeline

In order to explore the extent to which leveraging structured relations provides more meaningful results in extracting anatomical entities from clinical texts, a pipeline based on relations from UMLS is constructed. Specifically, the subset SNOMED CT is used, as it contains sufficient anatomical entities for the purposes of this study. The unique utility of UMLS lies in the relations it stores between biomedical entities. In Figure 3.1, the specific use-case is outlined. The disease entity "Acute Myocardial Infarction" points to the relevant anatomical structure "Heart Muscle" with the relation "has_finding_site".

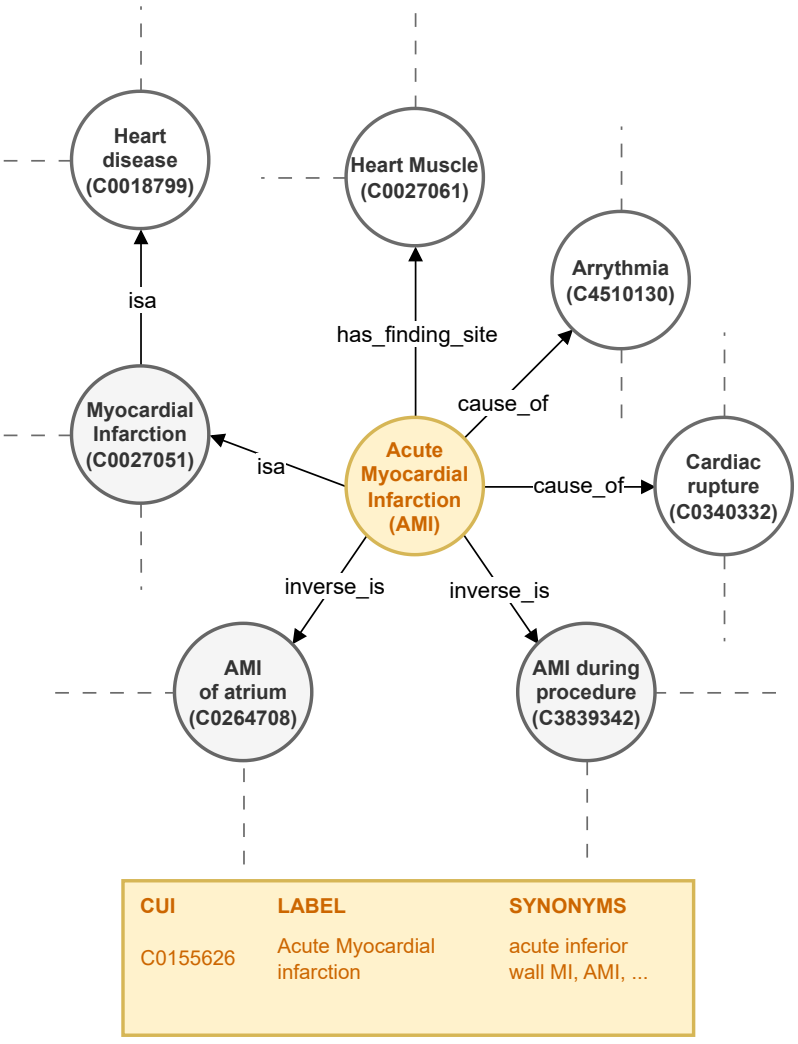


Figure 3.1: UMLS Relations

A combination of the tools explored in the previous sections are used to develop a robust pipeline that can extract implicit anatomical entities from explicit disease mentions. It follows the architecture outlined in Figure 3.2. The process begins with the extraction of disease mentions using the SciSpacy model en_ner_bc5cdr_md, which is trained specifically on disease mentions [13]. The extracted entities are then linked to Concept Unique Identifiers (CUIs) with SciSpacy's EntityLinker [48], which is a SpaCy component that maps text spans to standardized UMLS concepts. Unlike SciSpacys full NER pipeline with rule-based and statistical components built on the SpaCy library, this component uses string matching to link single entities to their corresponding UMLS CUIs. An example usage would be taking the string "Acute Myocardial Infarction" once detected by SciSpacy, and then producing the CUI "C0155626". This linking step is essential because the local database uses UMLS concept identifiers to define and store relationships between biomedical entities. By linking mentions to UMLS concepts, accurate and consistent retrieval of related concepts based on those pre-defined relations is made possible, bridging the gap between unstructured text and structured biomedical knowledge.

Once CUI's are obtained, a custom built database is queried to retrieve the anatomical entities related to the pathological process. A recursive lookup is included, where the parents of an entity are queried for their anatomical finding site in case the initial entity is not stored with one. Later, fuzzy string matching is applied to align the resulting anatomical concepts with predefined model label names in the 3D visualization environment, ensuring that each extracted concept can be accurately represented within the anatomical model.

In order to use meaningful relationships between disease entities and anatomical entities, a tailored local database is setup based on the latest 2025 SNOMED CT release, which is a subset of the UMLS release [49]. The database is comprised of two tables, built on files from the release. One table, MRCONSO, relates Concept Unique Identifiers (CUIs) to the english translation, and the other, MRREL, provides the relations between CUIs.

The reasoning for choosing SciSpacy for initial disease NER, is that this lightweight NER model is capable of identifying any entity type the aforementioned TUIs. This practical alternative is preferred to traditional NER models, which are limited to predefined entities, because this opens up the possibility to adapt the pipeline to any use case where the goal is to extract implicit biomedical information from explicit mentions. A simple adaption of the entity type to detect, as well as the relation to look up in the database will change the use case from the one explored in this thesis to any fit any task where implicit detection based on surface level mention is required. For example extracting specific cells implied by enzyme mentions, with the relation "affects", becomes possible through a simple change in the pipeline parameters. Other examples include relating biochemical processes mentioned in text to the tissues affected by these processes with the relation "occurs_in".

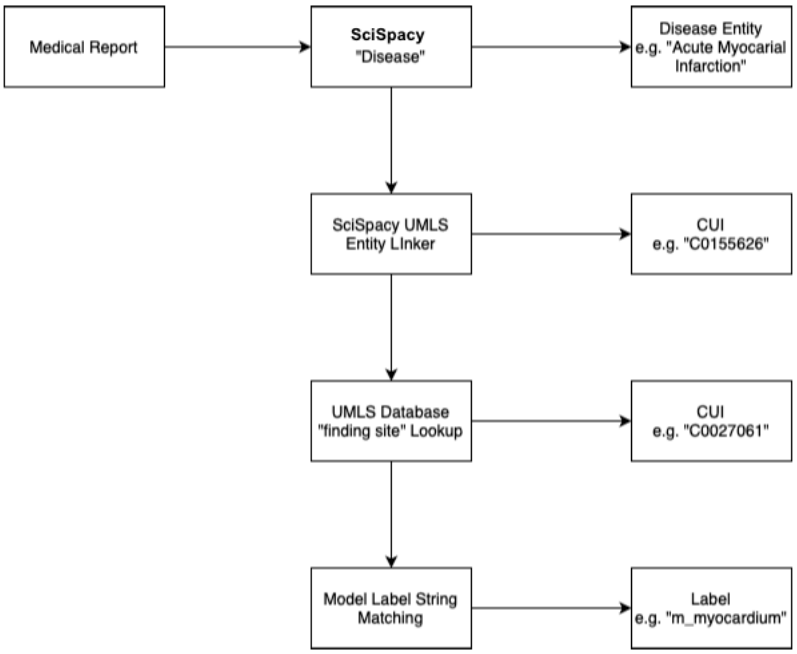


Figure 3.2: Pipeline Outline

3.4 Unity Integration

In parallel with entity extraction and anatomical concept resolution, this thesis explores optimal representations of visuals in a 3D Unity model, such that the selected anatomical structure is best visualized. Specifically, an analysis of available techniques and implementation methods is performed. A manual review of visuals is performed on a case-by-case basis, to tune several parameters that affect clarity and relevance. Finally, a contrast based evaluation of the diminishing returns of increasing inclusion radius of surrounding structures is performed, which is validated through comparison to the previously selected parameter.

Given a target anatomical region in the 3D model (e.g., "left ventricle"), the aim is to determine the most informative and clinically relevant camera position and orientation to highlight this region. The ideal visualization minimizes occlusion,

maximizes visibility, and provides appropriate anatomical context. Once a user inputs a clinical report to be visualized, the backend NER pipeline outlined in Figure 3.2 is triggered. To facilitate communication between the natural language processing (NLP) pipeline and the Unity-based 3D visualization environment, a RESTful API was developed using FastAPI, a modern Python web framework designed for high-performance applications. The schema for this process is outlined in Figure 3.3.

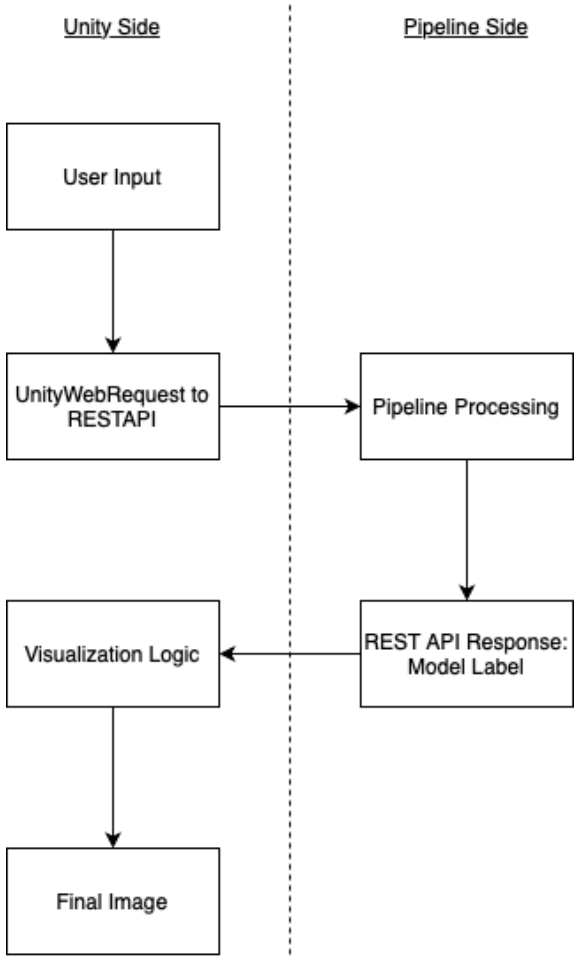


Figure 3.3: Pipeline Integration for Unity

3.4.1 Visualization logic

In order to explore which techniques optimize visualization, a new method is implemented. The main procedure is outlined by the pseudocode for the algorithm named ShowStructure in Algorithm 2. This function is designed to visually isolate and emphasize anatomical structures within a 3D model based on a search query. These matched objects are activated, highlighted, and brought into view by adjusting the camera focus. To provide contextual awareness, the method also activates nearby structures that are spatially close to the matched objects, unless they are excessively large or overlap in a way that could obscure the main focus. These contextual structures are shown in a dimmed and transparent state to distinguish

them from the primary focus. This approach allows users to clearly view the target anatomy while maintaining spatial awareness of its surroundings.

Algorithm 2 ShowStructure(queryName)

```
1: HideAllStructures
2: matched ← HighlightMatches(queryName)
3: if matched is empty then
4:     return
5: end if
6: FocusCameraOn(matched[0])
7: ShowContext(matched)
```

The script allows for adjustment of a "context distance" for which structures are still included in the visualization. This is combined with a further metric, "opacity" of surrounding structures, which controls their opacity, and a control for dimming the surrounding structures. These three metrics are tested to find optimal values for visual representation.

The algorithm begins by hiding all elements and restoring their original materials to ensure a clean slate with the sub process HideAllStructures, outlined in Algorithm 3. It then searches through all renderers in the model for the anatomical entity or entities produced by the NER pipeline. The queried anatomical structures is isolated using Algorithm 4. The camera is focused on the structure based on its position in the scene, and the size of the structure with FocusCameraOn, outlined in Algorithm 5. Spatial awareness through context visualization in (Algorithm 6) is maintained, by iterating through all renderer components under a root anatomical object (anatomyRoot) to determine which surrounding context objects should be visually displayed alongside a set of matched objects. For each renderer (representing a potential context object), it skips rendering if the object is either already in matchedObjects or is unusually large in size (likely to be irrelevant or obtrusive). It then compares each candidate context object against each matched object by computing spatial relationships. Specifically, whether the objects are close (using the distance between bounding boxes) and whether the context object is disproportionately larger. If the bounding boxes overlap and the context object is much larger, it is skipped to avoid visual clutter. Otherwise, if the context object is within a predefined distance (contextDistance), the algorithm activates it and its parent hierarchy, enables its renderer, and dims it visually to provide context without overpowering the matched objects, outlined in Algorithm 7.

3.4.2 User Interface

The user interface includes two components, an input field for medical reports, as well as a button for triggering the back end analysis and front end visualization. The input field accepts unstructured clinical text and is configured to handle both complete impressions and single-sentence findings. Once the user submits the input, the backend processes the report by extracting relevant pathological and anatomical entities using the pipeline described in previous sections. The trigger button

Algorithm 3 HideAllStructures

```
for all transform t in children of anatomyRoot do
    if t ≠ anatomyRoot then
        deactivate t's GameObject
    end if
end for
for all (object, material) in stored original materials do
    if object has Renderer then
        set material to original material
    end if
end for
clear stored original materials
```

Algorithm 4 Highlight(go)

```
if go has no Renderer then
    return
end if
save current material in _originalMats
m ← a new copy of the current material
enable the _EMISSION keyword
set emission color to yellow × 2
assign material m to the Renderer
```

Algorithm 5 FocusCameraOn(target)

```
if mainCamera is null or target has no Renderer then
    return
end if
bounds ← target's Renderer bounds
center ← bounds center
dir ← normalized vector (left + 0.5 × up)
dist ← magnitude of bounds extents × 3
newPosition ← center + (dir × dist)
set mainCamera position to newPosition
rotate mainCamera to look at center
```

Algorithm 6 ShowContext(matchedObjects)

```
allRenderers ← all Renderer components in anatomyRoot
for all contextRenderer in allRenderers do
    ctxObject ← contextRenderer's GameObject
    if ctxObject in matchedObjects then
        continue
    end if
    if contextRenderer's bounds size is very large then
        continue
    end if
    for all mainObject in matchedObjects do
        mainBounds ← bounds of mainObject
        ctxBounds ← bounds of contextRenderer
        closestPoint ← point on mainBounds closest to ctxBounds center
        gap ← distance between closestPoint and ctxBounds center
        boxesOverlap ← (gap ≈ 0)
        sizeRatio ← ctxBounds size / mainBounds size
        if boxesOverlap and sizeRatio > 2.5 then
            continue
        end if
        if gap ≤ contextDistance then
            activate ctxObject and its parents
            enable contextRenderer
            visually dim ctxObject
            break
        end if
    end for
end for
```

Algorithm 7 Dim(go)

```
if go has no Renderer then
    return
end if
if go not in _originalMats then
    save current material in _originalMats
end if
mat ← a new copy of the current material
if mat has property _Surface then
    set _Surface to 1 (transparent)
end if
if mat has property _Mode then
    set _Mode to 3 (fade mode)
end if
set blend and render properties for alpha transparency:
  SrcBlend = SrcAlpha, DstBlend = OneMinusSrcAlpha
  ZWrite = 0
  disable _ALPHATEST_ON, enable _ALPHABLEND_ON
  set render queue to 3000
dim RGB color by 50% and set alpha to DIM_ALPHA
assign modified material to the Renderer
```

serves as a single entry point to execute the entire workflow, from Named Entity Recognition (NER) and entity linking to the rendering of anatomical structures in the 3D viewer. The interface is implemented within Unity and designed to present an all-in-one method of going from report to visualization.

3.5 Evaluation Design

The evaluation strategy aimed to determine whether the proposed pathology-driven pipeline improves the extraction quality of clinically relevant anatomical sites compared to three baseline approaches: (1) surface-level anatomical entity extraction, (2) LLaMA2 prompting for underlying entities, and (3) transformer-based anatomical entity extraction. This section details the evaluation metrics, validation of the LLM-based assessment method, and statistical analyses employed.

3.5.1 NLP Extraction Evaluation

3.5.1.1 Annotation Quality Scoring

Annotation quality was assessed using two independent large language models (LLMs) acting as evaluators: LLaMA3 and Mistral. Each LLM was prompted to assign a score based on the relevance of the anatomical entity produced by each method to the medical report on a five-point ordinal scale (1 = incorrect or irrelevant, 5 = highly accurate and clinically appropriate). This scale was selected for its inter-

pretability and ability to capture degrees of correctness rather than binary outcomes. The evaluation was conducted on 1,100 radiology reports processed by each of the four extraction methods. This design ensured that every report was assessed across all approaches by both LLM judges, enabling statistical comparisons. The reason for choosing this method is that human evaluation is costly and time-consuming, and LLMs can model expert reasoning when properly prompted [50].

LLM-As-A-Judge Prompt:
"You will be given a clinical report impression, as well as a prediction for the most relevant anatomical site mentioned or implied. Your task is to rate the prediction's relevance on a scale from 1-5.

Evaluation Steps:

Read the prediction and compare it to the report impression. Check if the prediction accurately describes the anatomical finding site of the main pathology discussed. Assign a relevance score on a scale of 1 to 5, where 1 is the lowest, and 5 is the highest. In case the report states that no pathologies have been found, the prediction "none" should receive a score of 5, whereas an anatomical finding site should receive 1.

Report Impression: {impression}

Prediction: {prediction}

Respond with only the score.

Answer:"

3.5.1.2 Validation of LLM-Based Evaluation

Given the novelty of using LLMs as evaluators, a validation experiment was conducted to assess their reliability. A subset of 100 radiology reports was selected, for which two annotation types were generated: (a) a gold-standard annotation, and (b) a nonsensical annotation created by randomly generating unrelated words. Both annotations were scored by LLaMA3 [51] and Mistral [52] using the same five-point scale. Discrimination performance was measured using three metrics. MannWhitney U Test: To determine whether LLM-assigned scores for gold-standard annotations were significantly higher than for nonsensical annotations. Effect Size (Cohens d): To quantify the magnitude of the difference in scores. ROC Analysis and AUC: Treating gold-standard annotations as positive cases and nonsense as negative, the area under the ROC curve (AUC) was computed as an indicator of classification reliability. This validation step aims to ensure that the automated scoring framework provides a robust proxy for human evaluation.

3.5.1.3 Statistical Analysis for Main Comparison

Since quality scores are ordinal and each report was evaluated under all four conditions, non-parametric tests were selected. The Friedman test was used to detect overall differences in annotation quality across the four methods for each LLM judge. Effect size for this global test was reported using Kendalls W, where values of 0.1, 0.3, and 0.5 indicate small, moderate, and large effects, respectively. Where the Friedman test indicated significance, pairwise differences were assessed using the Wilcoxon signed-rank test with Bonferroni correction to control for family-wise error rate. This post-hoc analysis identified which methods significantly outperformed others. All analyses were conducted in Python using the SciPy and scikit-learn libraries.

3.5.2 Unity Integration Evaluation

3.5.3 Parameter Sweep Design

To investigate how the inclusion of surrounding anatomical structures affects target visibility in 3D medical visualizations, a parameter sweep was implemented in Unity. The sweep varied two key parameters: ctxDist, the radius within which surrounding anatomical structures are included, and dimAlpha, the alpha transparency level applied to the context. The ctxDist parameter varied from 0.01 m to 0.1m across 50 steps, while dimAlpha varied from 0.0 (fully transparent) to 1.0 (fully opaque) across 5 levels, resulting in 250 unique visualization conditions per organ. This was repeated for 5 major organs (lung, liver, stomach, prostate and colon) for a total of 1250 renderings.

Each trial rendered a target anatomical structure while including nearby structures as context. Contextual structures were displayed with variable opacity (dimAlpha) to simulate different visual emphasis. This approach allowed the exploration of how varying levels of anatomical context influence the perceptual salience of the target.

3.5.4 Rendering and Masking Pipeline

Each trial rendered two distinct visual passes using Unitys camera and shader system: one for the highlighted target structure, and one for the contextual anatomy. Custom binary masks were generated for both passes using replacement shaders. These masks allowed the isolation of luminance contributions from the target and the context, respectively.

Color buffers and binary masks were captured from the GPU using RenderTexture and read into CPU memory using the ReadPixels function. From this, pixel-wise luminance was computed using standard weights for RGB conversion. The results were then stored in a CSV file.

3.5.5 Visibility Metrics

To quantify the visual prominence of a target structure within its anatomical surroundings, luminance-based metrics were computed from GPU-rendered frames. For each trial, two binary masks were generated to isolate the target and the surrounding structures using separate rendering passes. Pixel luminance was computed using the Rec. 709 formula for perceived brightness [53]:

$$L = 0.2126 \cdot R + 0.7152 \cdot G + 0.0722 \cdot B$$

From this, two scalar values were derived:

- L_h Mean luminance of the target structure (highlight mask).
- L_c Mean luminance of the anatomical context (context mask).

Additionally, the number of structures within the inclusion radius was recorded as the **context count** N_c.

A contrast score was computed to quantify the luminance-based distinguishability of the target structure relative to its background:

$$\text{Contrast} = \frac{\max(L_h, L_c) + \varepsilon}{\min(L_h, L_c) + \varepsilon}, \quad \text{with} \quad \varepsilon = 0.05$$

This ratio-based formulation is invariant to which region is brighter and avoids the undefined behavior of standard contrast metrics when luminance values approach zero. It is designed to reflect relative perceptual salience: a value close to 1 implies low visual separation, whereas higher values indicate a more pronounced distinction between target and context. While high contrast supports target visibility, excessive anatomical context may introduce visual clutter, which can degrade performance. This phenomenon has been documented in both perception research and applied fields like interface design and radiology [54]. As such, the visibility score used in this study incorporates a clutter penalty that accounts for both the luminance and the quantity of surrounding structures. The contrast metric above forms the foundation of this composite score, serving as a perceptually grounded estimate of initial visibility before contextual interference is considered. ϵ was chosen at 0.05 empirically to suppress noise from dark regions, as it is relatively small to typical luminance values.

3.5.6 Visibility Score

To reflect how anatomical context affects perceptual salience, a visibility score was defined as:

$$\text{Visibility Score} = \frac{\text{Contrast}}{L_c \cdot N_c}$$

This score penalizes visibility when either the average background brightness (L_c) or the amount of contextual clutter (N_c) increases. The formulation is intended to represent perceptual load: a structure becomes harder to perceive not just when contrast drops, but also when it is embedded in a dense and bright surrounding environment. As L_c increases, the background becomes more visually competitive; as N_c increases, the scene becomes more semantically crowded.

By plotting visibility score against context radius ($ctxDist$), the point of diminishing returns can be estimated using knee-point detection or nonlinear curve fitting. A rapid initial drop followed by a plateau is considered indicative of perceptual saturation, where added anatomical context no longer improves spatial comprehension but instead reduces visual clarity.

The previously outlined manual analysis resulted in a context distance parameter of 2 cm, which was chosen as a reference value to compare this approach to. While this baseline does not represent a ground-truth optimum derived from user studies, it serves as a perceptual reference. This value aligns with real-world radiological practice, where excessive context can obscure key structures, and minimal context may hinder spatial understanding [55]. Therefore, the 2 cm value offers a reasonable threshold to test whether automatic detection methods yield comparable results.

3.5.7 Validation of Contrast-Based Visibility Score

In order to validate the contrast based evaluation of diminishing returns with increasing inclusion radius of surrounding structures, the threshold beyond which additional context yields minimal perceptual benefit was calculated based on the contrast data, in order to compare it to the manual selection, to analyze whether it aligns with human perception. A fixed baseline of 2 cm was used as a reference point for optimal anatomical context inclusion. The 2 cm value provides a neutral benchmark to assess whether algorithmic methods can yield plausible and perceptually meaningful context distances. Rather than acting as a definitive ground truth, it functions as a reference for comparative validation.

3.5.7.1 Kneedle Algorithm

The Kneedle algorithm identifies the point of maximum curvature in a normalized score-distance curve by measuring the perpendicular distance between each point on the curve and the diagonal line that connects the start and end points.

Let the normalized visibility curve be defined as:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}, \quad y' = \frac{y - \min(y)}{\max(y) - \min(y)}$$

The baseline line from (x_1, y_1) to (x_2, y_2) is defined as a vector:

$$\vec{v} = \begin{bmatrix} x_2 - x_1 \\ y_2 - y_1 \end{bmatrix}, \quad \text{normalized as } \vec{u} = \frac{\vec{v}}{\|\vec{v}\|}$$

Then, for each point (x_i', y_i'), the perpendicular (orthogonal) distance to the line is:

$$d_i = \left\| \begin{bmatrix} x_i' - x_1 \\ y_i' - y_1 \end{bmatrix} - \left[\left(\begin{bmatrix} x_i' - x_1 \\ y_i' - y_1 \end{bmatrix} \cdot \vec{u} \right) \vec{u} \right] \right\|$$

The knee point is selected as the x-value corresponding to the maximum d_i.

3.5.7.2 L-method

The L-method fits two linear segments to a score-distance curve and finds the splitting index k that minimizes the combined residual error of both linear fits. For a candidate index k:

Line 1 is fit to the first segment: $y = m_1 x + b_1$ for $x \leq x_k$, and line 2 is fit to the second segment: $y = m_2 x + b_2$ for $x > x_k$

Finally, the sum of squared errors (SSE) is computed for both:

$$\text{SSE}_1 = \sum_{i=1}^{k} (y_i - (m_1 x_i + b_1))^2, \quad \text{SSE}_2 = \sum_{i=k+1}^{n} (y_i - (m_2 x_i + b_2))^2$$

The optimal knee point is the value of x_k for which the total SSE is minimized:

$$x^* = \arg\min_k (\text{SSE}_1 + \text{SSE}_2)$$

This method is particularly effective when the curve transitions from a steep increase to a plateau, as it identifies the structural break point in the curve.

To assess how closely the automatically determined knees align with the perceptual 2 cm baseline, two statistical tests were applied:

- A one-sample t-test compared the mean knee distances detected by Kneedle to the fixed 2 cm baseline.
- A paired t-test compared the distances obtained by Kneedle and the L-method across organs.

The mean absolute deviation from the baseline was also computed for both methods.

4

Results

4.1 Extraction Quality Comparison

Table4.1 summarizes the mean, standard deviation, and median scores assigned by LLaMA3 and Mistral to each extraction approach. The proposed pipeline achieved the highest average quality score across both evaluators (4.41 for LLaMA3, 3.51 for Mistral), outperforming surface-level extraction and transformer-based models. Although LLaMA3 and Mistral differed in strictness, both ranked the proposed method among the best-performing systems.

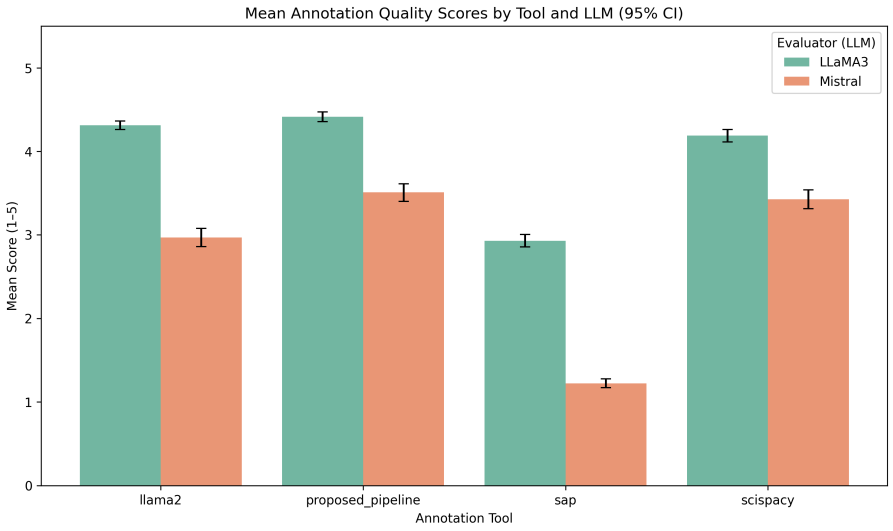


Figure 4.1: LLM-As-A-Judge Results

Statistical analysis using the Friedman test indicated significant differences in annotation quality across the four methods for both LLaMA3 ($\chi^2 = 981.04$, $p < 0.001$) and Mistral ($\chi^2 = 923.43$, $p < 0.001$), with moderate effect sizes (Kendalls W = 0.297 and 0.280, respectively). Post-hoc Wilcoxon signed-rank tests with Bonferroni correction confirmed that the proposed pipeline significantly outperformed sap ($p < 0.001$ for LLaMA3, $p < 0.001$ for Mistral) and showed meaningful improvements over LLaMA2 prompting ($p = 0.004$ for LLaMA3, $p <$ 4.07e-12 for Mistral). Against scispaCy, the proposed method scored higher with LLaMA3 ($p < 0.001$) but did not differ significantly under Mistral ($p = 1.0$).

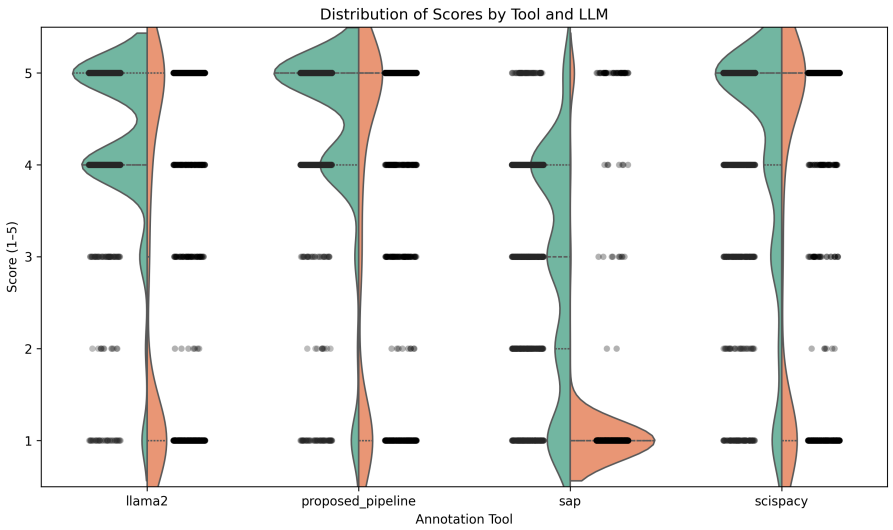


Figure 4.2: LLM-As-A-Judge Violin Plot

LLM	Tool	Mean Score	Std Dev	Median
LLaMA3	LLaMA2	4.31	0.88	4.0
LLaMA3	Proposed Pipeline	4.41	1.00	5.0
LLaMA3	SAP	2.93	1.27	3.0
LLaMA3	scispaCy	4.19	1.27	5.0
Mistral	LLaMA2	2.97	1.88	3.0
Mistral	Proposed Pipeline	3.51	1.78	5.0
Mistral	SAP	1.22	0.89	1.0
Mistral	scispaCy	3.43	1.88	5.0

Table 4.1: Mean, standard deviation, and median quality scores assigned by LLaMA3 and Mistral to each extraction method across 1,100 reports.

These findings suggest that incorporating pathologyanatomy inference via SNOMED CT substantially improves the relevance of extracted anatomical sites compared to transformer based approaches and general-purpose prompting. However, differences against rule-based extraction methods such as ScispaCy are less prominent under certain evaluation conditions, indicating that contextual sensitivity remains a challenge.

LLM	Comparison	p-value	Significant
LLaMA3	LLaMA2 vs Proposed Pipeline	$p < 0.001$	Yes
LLaMA3	LLaMA2 vs SAP	$p < 0.001$	Yes
LLaMA3	LLaMA2 vs scispaCy	0.160	No
LLaMA3	Proposed Pipeline vs SAP	$p < 0.001$	Yes
LLaMA3	Proposed Pipeline vs scispaCy	$p < 0.001$	Yes
LLaMA3	SAP vs scispaCy	$p < 0.001$	Yes
Mistral	LLaMA2 vs Proposed Pipeline	$p < 0.001$	Yes
Mistral	LLaMA2 vs SAP	$p < 0.001$	Yes
Mistral	LLaMA2 vs scispaCy	$p < 0.001$	Yes
Mistral	Proposed Pipeline vs SAP	$p < 0.001$	Yes
Mistral	Proposed Pipeline vs scispaCy	1.00e+00	No
Mistral	SAP vs scispaCy	$p < 0.001$	Yes

Table 4.2: Pairwise Wilcoxon signed-rank test results comparing annotation quality scores across extraction methods, with Bonferroni correction. Significance threshold: $\alpha = 0.05$.

LLM	Friedman χ^2	Kendall's W
LLaMA3	981.04	0.2973
Mistral	923.43	0.2798

Table 4.3: Friedman test results for quality score differences across the four extraction methods. Kendalls W indicates effect size.

4.2 Reliability of LLM-based Evaluation

To assess the validity of LLM-based scoring as a proxy for human evaluation, a discrimination test between gold-standard annotations and nonsensical annotations was conducted for a subset of 100 reports. Both LLaMA3 and Mistral demonstrated strong reliability in distinguishing annotation quality. As shown in Table 4.4, the mean score for gold annotations was substantially higher than for nonsense (4.61 vs 1.67 for LLaMA3; 4.06 vs 1.00 for Mistral). MannWhitney U tests confirmed that these differences were highly significant ($p < 0.001$), with extremely large effect sizes (Cohens $d > 2.8$).

Table 4.4: LLM Reliability Results

LLM	Mean Gold	Mean Nonsense	p-value	Cohen's d	AUC
LLaMA3	4.610	1.670	0.000	3.292	0.962
Mistral	4.060	1.000	0.000	2.854	0.915

ROC analysis further supports the robustness of LLM judgments. LLaMA3 achieved an AUC of 0.962, and Mistral 0.915, both indicating excellent discriminative performance (Figure 4.3).

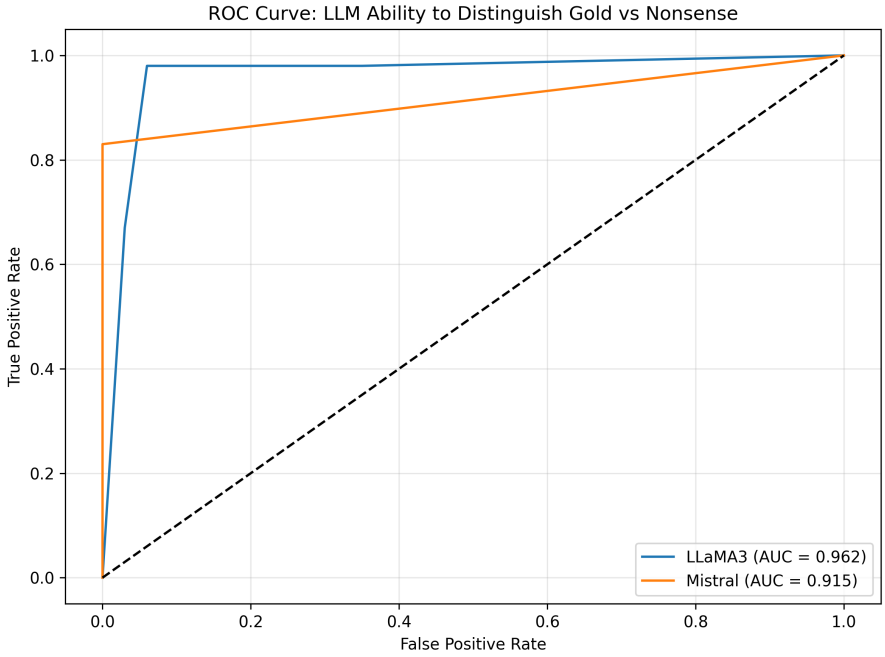


Figure 4.3: ROC Curves Mistral and LLaMA 3

These results suggest that LLM-based evaluation is a reliable and scalable alternative to human judgment for this task. While LLMs may exhibit evaluator bias or variability, the strong separation between high- and low-quality annotations indicates that they are suitable for benchmarking methods when human evaluation is impractical. The results demonstrate that the proposed pathology-driven approach

offers a measurable improvement in extracting clinically relevant anatomical sites. By leveraging SNOMED CT for semantic linkage, the pipeline outperforms surface-level extraction and general LLM prompting strategies, achieving the highest quality ratings under both evaluators. This indicates that grounding anatomical inference in pathology information provides a meaningful enhancement over simpler heuristics or single-step extraction models. Importantly, validating LLM-based evaluation shows the credibility of the comparative analysis. The ability of LLaMA3 and Mistral to accurately distinguish gold-standard from nonsense annotations (AUC > 0.9) suggests that these models can serve as cost-effective, reproducible evaluators, reducing the need for extensive human annotation in future studies. Nonetheless, some limitations remain. Differences in scoring behavior across LLMs highlight the potential for evaluator bias, and reliance on SNOMED CT assumes accurate mapping, which may fail in ambiguous or rare cases. Future work could integrate probabilistic reasoning or ontology-driven inference to further enhance robustness.

4.3 Differences in NER Techniques

The following example report on a patient diagnosed with pneumonia, which does not explicitly mention anatomical finding sites such as the lungs, highlights the key differences in the approaches outlined in this thesis. Specifically, an analysis of the way context is handled by different techniques is explored, and similar results were seen in all cases tested.

Example Clinical Report:
"The patient presented with a several-day history of fever, productive cough, chills, and generalized fatigue. On examination, vital signs revealed an elevated temperature and increased respiratory rate, with oxygen saturation slightly below normal limits. Auscultation detected abnormal breath sounds localized to one side of the chest, accompanied by dullness to percussion. A chest radiograph confirmed a localized infiltrative process. The patient was diagnosed with pneumonia and initiated on empiric antibiotic therapy, with supportive care provided as needed. Clinical improvement was observed within 48 hours of treatment initiation."

SciSpacy's en_core_sci_sm NER model was tested on its capabilities to extract relevant anatomical entities, and the output of this sample report highlights the key weakness of this off-the-shelf NER technique: it relies on surface level mentions to detect entities.

SciSpacy's Detected Entities:
patient, several-day history, fever, productive cough, chills, generalized fatigue, examination, revealed, elevated, ature, increased, respiratory rate, oxygen saturation, Auscultation, detected, abnormal, breath sounds, localized, dullness,

percussion, chest radiograph, ized, infiltrative process, patient, diagnosed, pneumonia, initiated, empiric, antibiotic therapy, supportive care, Clinical, improvement, treatment.

While the main issue is listed ("pneumonia"), there is no link to anatomical context, as seen by the fact that "lung" was not returned. The anatomical entity that survived the filter was "chest", since this was mentioned explicitly, however this is not sufficiently detailed. Given that anatomical entities are most often implied through pathological context and not mentioned explicitly, this technique is not suitable to achieve the desired results for visualization in the anatomical model.

The output produced by SapBERT made very little sense in this case. A mixture of muscles and tissues were listed, none of which significantly relate to the underlying disease. In other cases, some related structures were produced, such as arteries supplying the target structure, but rarely was the main entity detected.

SapBERT's Detected Entities:

- Posterolateral Med (cosine=0.408)
- Antitragicus (cosine=0.339)
- Temporoparietal Fascia (cosine=0.339)
- Fascia Mobile Wad Compartment (cosine=0.329)
- Lateral Temporal-Cheek Fat Pad (cosine=0.327)

LLaMA was prompted to provide the single most relevant anatomical entity in the report, and to only respond with the entity. It responded with "Lung", which is correct. Similar results were achieved in other cases, where LLaMA was capable of extracting implicit anatomical concepts, yet the granularity was limited. LLaMA was trained on 2 trillion tokens [15], which is an impressive feat. However, most of this pretraining data is not specific biomedical text, therefore a detailed contextual understanding of relationships between pathology and anatomy is unlikely to be seen for the vast world of clinical reports.

4.3.1 Novel Proposed UMLS-Enhanced Pipeline

The proposed pipeline recognized "Lung" as the correct entity in the previous report. In other reports, the UMLS pipeline shows its ability to extract underlying anatomy with more nuance than the baseline techniques. For example, in the sentence: "CT reveals a large mass in the pancreatic head causing biliary obstruction", the pipeline outputs the biliary duct, while LLaMa got distracted by the mention of the pancreatic head, and output "pancreas". SciSpacy recognized "pancreatic head" as a mention, and SapBERT output structures related to the pancreas, such as the greater pancreatic artery. In Figure 4.4, the anatomical relations between the pancreas, bile duct and liver can be observed. The pathology leading to the biliary obstruction

lies in the enlargement of the pancreatic head, however the main issue lies in the biliary duct itself [56]. This is an example case where the proposed pipeline shines, as baseline techniques produce the pancreatic head, but the UMLS-enhanced model recognizes the biliary duct as the main problematic anatomical entity. A biliary obstruction may lead to acute issues such as jaundice and cholestasis, which a doctor may explain and treat immediately [57]. The enlargement of the pancreatic head also requires attention, but the immediate threat lies in the biliary duct obstruction.

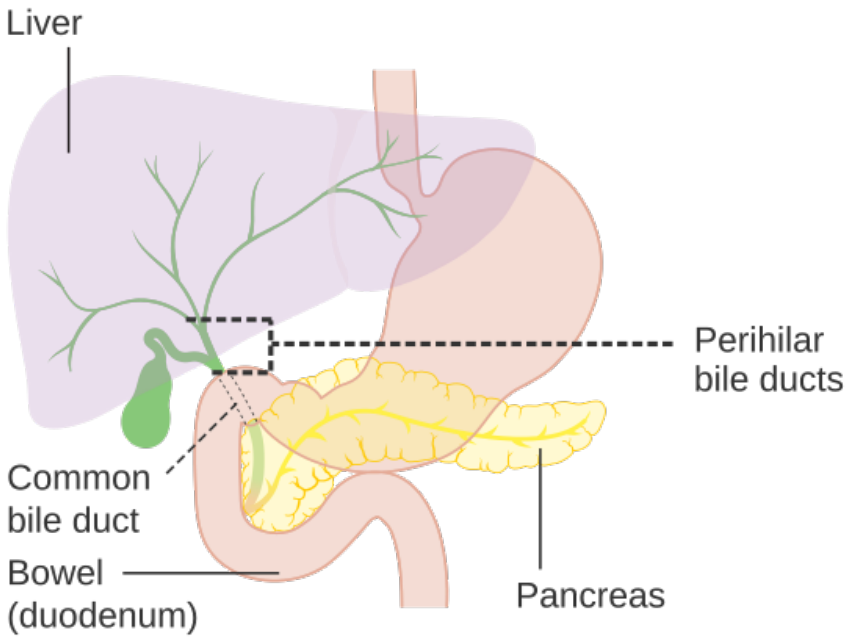


Figure 4.4: Anatomical Relationships of Gastrointestinal Structures. Image by Cancer Research UK

4.4 Unity Visualization

For the visual analysis of the entity representation, examples will be based on the previously mentioned report detailing a patient's pneumonia diagnosis, as well as a report on a patient with carpal tunnel syndrome, where the affected structure is the median nerve. The UMLS-based pipeline correctly identified the most pathological process in the first report as pneumonia, and output the model label "Lung_R". In the second report, the label "Median_Nerve_R", was output. A manual analysis for the key parameters relating to visual representation was performed for several cases, and an exemplary flow is described in the following sections. This is followed by statistical analysis for the context distance parameter, which was performed on 5 organs with 250 renderings each to analyze the diminishing returns experienced with added structures.

4.4.1 Entity Representation

The algorithm for panning the camera and zooming as outlined in the methodology successfully placed the camera at positions where visualization is optimal. Through testing with several structures, the optimal value for the zoom was determined. The algorithm moves the camera back from its original position to a value multiplied by the object's half-diagonal, ensuring that the entire object will fit in view. The following examples show that for a value of 3, the structure is central in the frame, and clearly represented.

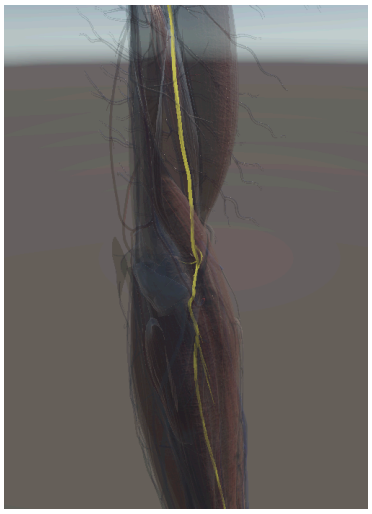


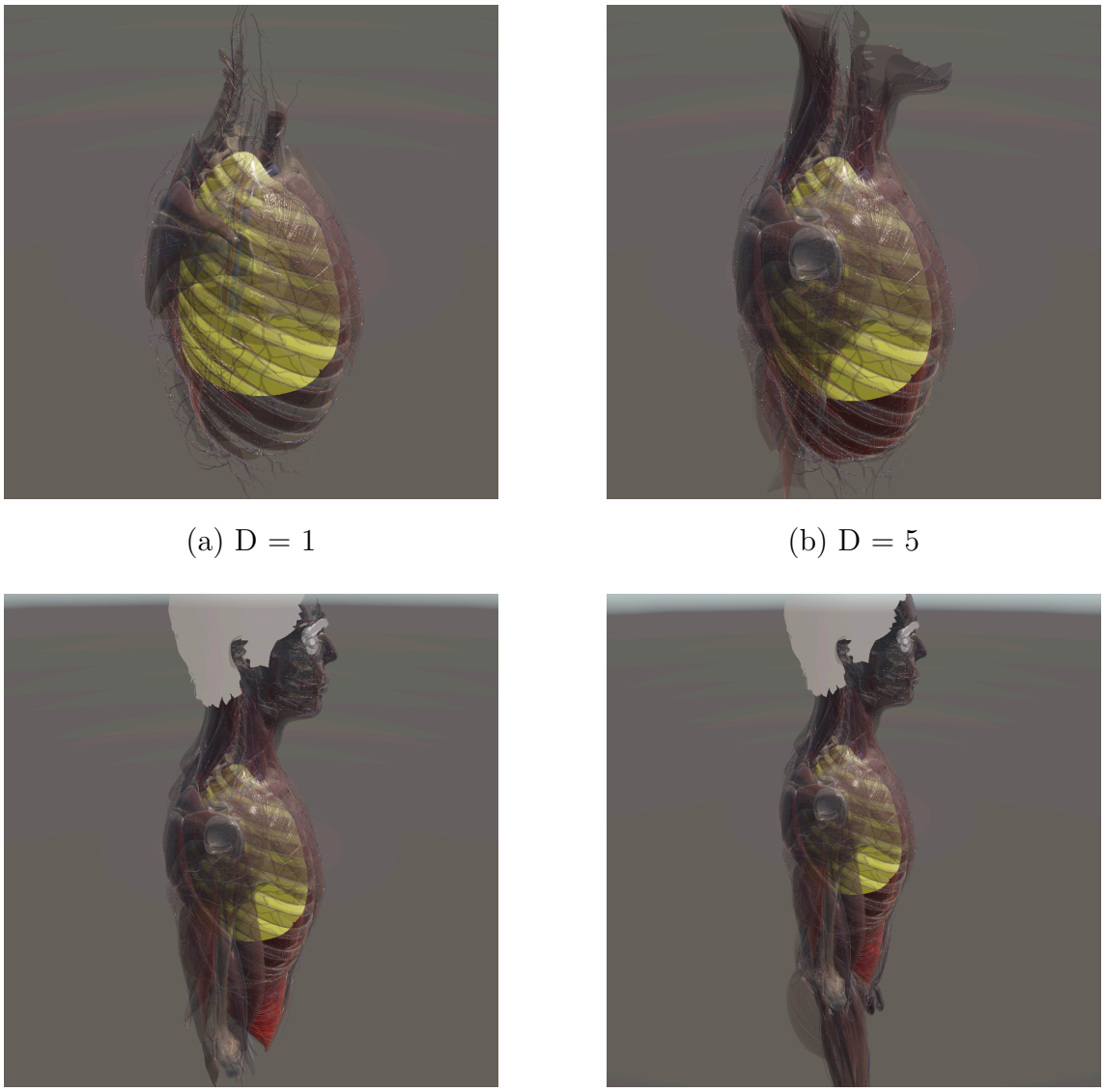
Figure 4.5: Zoom Value = 1.0

Figure 4.6: Zoom Value = 3.0

Figure 4.7: Zoom Value = 5.0

A context distance parameter D is introduced into the highlighting pipeline. During each render pass, any mesh whose minimum center-to-surface distance from the target anatomy is less than or equal to D is retained in the scene, while others are omitted. By tuning D between 0 (no context) and 50 units (very broad context), the trade-off between isolation and situational awareness in the 3D view can be controlled. Figure 4.8 showcases some sample D values, the ideal value was determined to be 2. The reason for choosing center-to-surface rather than surface-to-surface or center-to-center was that results showed better consistency in visualization over the other methods. This is likely due to the fact that larger objects such as the lungs would have more structures included when taking surface-to-surface as the gap measurement, thereby causing more visual clutter.

The adjustable metric for dimming colors of surrounding structure is an effective tool for reducing distractions and maintaining the focus on the relevant entity. Selecting an appropriate value required testing with several structures. With surrounding objects at full dimming, surrounding structures lose all color, maximizing the focus on the main target entity, but reducing context as seen in Figure 4.9, where all related structures are dull and grey. Without dimming, showcased in Figure 4.11, the target structure becomes less prominent, reducing focus. In Figure 4.10, dimming is at 0.5, which was deemed as the ideal tradeoff after testing with several structures.



(a) D = 1

(b) D = 5

(c) D = 20

(d) D = 50

Figure 4.8: Comparison of Various Context Distance Values (cm)

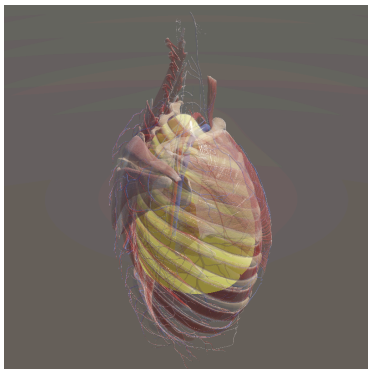


Figure 4.9: Dimming = 1.0

Figure 4.10: Dimming = 0.5

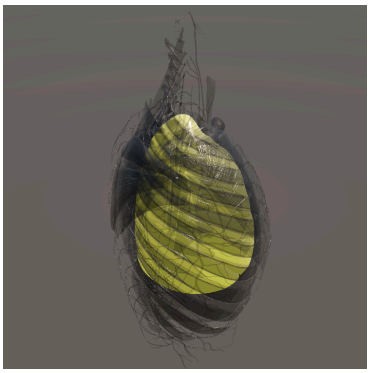


Figure 4.11: Dimming = 0.0

The final adjustable metric affects the opacity of surrounding objects. This had the most drastic effect on visualization. With surrounding objects at full opacity, underlying target structures remain hidden as seen in figure 4.12, where the lungs are obstructed by the ribs and muscles. With objects at zero opacity, showcased in figure 4.14, only the target structure itself remains visible, removing valuable context for the patient. In figure 4.13, opacity is at 0.5, highlighting that a good tradeoff is required. With iterative testing of several structures, the ideal opacity was determined to be 0.4.

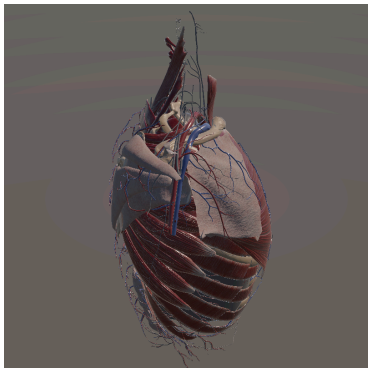


Figure 4.12: Opacity = 1.0

Figure 4.13: Opacity = 0.5

Figure 4.14: Opacity = 0.0

The final optimal representation with the manually chosen parameters (zoom 3, context distance 2, opacity 0.4, dimming 0.5) produces Figure 4.15. This is an appropriate visualization for a pneumonia patient, who may be confused about the various symptoms they are experiencing, such as fever and weakness, to understand the root location of their illness.

Unfortunately, a female version of the anatomical model was not available for this study. Future implementations should provide visualizations for both sexes, as sex specific anatomy plays a crucial role in medicine, and tools should not be limited to only one sex. An exemplary image to guide future implementations can be seen in figure 4.16. This representation of a uterus with surrounding structures is an idea of what can be expected from visualizations of medical reports on pathologies of the female reproductive organs.

Figure 4.15: Visual Representation for Pneumonia Patient

4.4.2 Visibility as a Function of Anatomical Radius

Visibility scores were computed across a dense sweep of inclusion radii (*ctxDist*) and context opacity levels (*dimAlpha*) for five anatomical structures: liver, prostate, stomach, colon, and lung, and 250 renderings per organ. For each organ, a characteristic curve was observed in which visibility decreased rapidly with small increases in *ctxDist* and then gradually stabilized.

4.4.3 Evidence for Diminishing Returns

To test whether visibility reduction followed a pattern of diminishing returns, the visibility curve for each organ was modeled using both a linear regression and an exponential decay function of the form:

$$y = a \cdot e^{-bx} + c$$

Both models were fit to the visibility scores as a function of *ctxDist*. Model fit was evaluated using the coefficient of determination (R^2), and model preference was determined using a likelihood ratio test (LRT) comparing the log-likelihoods of the exponential and linear fits.

Table 4.5 summarizes the R^2 values and log-likelihoods for both models, alongside the LRT statistic and corresponding p-value. A significant LRT ($p < 0.05$) indicates that the exponential model provides a statistically better fit than the linear alternative, consistent with a nonlinear decline in visibility.

table

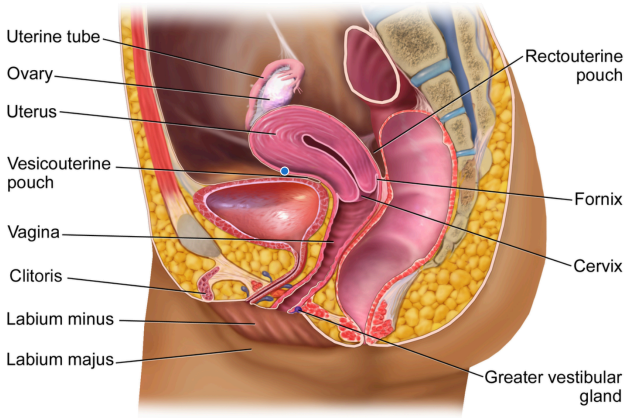


Figure 4.16: Visual Representation for Female Patient with Pathological Findings in Reproductive Organs licensed under CC BY 3.0

Table 4.5: Model comparison for visibility score vs. context radius

Organ	R^2_{linear}	R^2_{exp}	$\log L_{\text{linear}}$	$\log L_{\text{exp}}$	LRT χ^2	p-value
Liver	0.6743	0.9836	470.577	844.333	747.511	< 0.001
Prostate	0.4592	0.9490	485.848	781.036	590.376	< 0.001
Stomach	0.2857	0.9557	-2.097	345.521	695.237	< 0.001
Colon	0.5443	0.9932	474.003	999.243	1050.479	< 0.001
Lung	0.4848	0.9939	272.513	826.429	1107.831	< 0.001

As shown, all organs displayed a strong preference for the exponential model, with highly significant p-values and marked increases in R^2. This suggests that visibility declines rapidly with context inclusion, but that the perceptual cost tapers off, meaning diminishing returns are present.

4.4.4 Knee Point Analysis

In order to validate the contrast based evaluation of diminishing returns with increasing inclusion radius of surrounding structures, the threshold beyond which additional context yields minimal perceptual benefit was calculated based on the contrast data, in order to compare it to the manual selection, to analyze whether it aligns with human perception. Knee points were detected using two algorithms: the Kneedle method (based on maximum perpendicular distance to a linear baseline) and the L-method (based on piecewise linear error minimization). These points mark the onset of diminishing returns and suggest an optimal cutoff for spatial inclusion.

Across organs, the knee point typically occurred between 1.5—3.5 cm, suggesting that contextual inclusion beyond this radius contributes relatively little to perceptual clarity under the tested rendering conditions. There is however variation in the point of diminishing returns, where smaller organs experience sooner, and larger organs later diminishing returns. The manual selection for optimal context distance falls in the range of the knee points, showcasing that the analysis method sufficiently models visibility for the purposes of this investigation.

Table 4.6: Context distance (in meters) selected by each method for each organ.

Organ	Manual (2 cm)	L-method	Kneedle Knee
Liver	0.0200	0.0265	0.0302
Prostate	0.0200	0.0173	0.0210
Stomach	0.0200	0.0173	0.0210
Colon	0.0200	0.0247	0.0338
Lung	0.0200	0.0247	0.0284

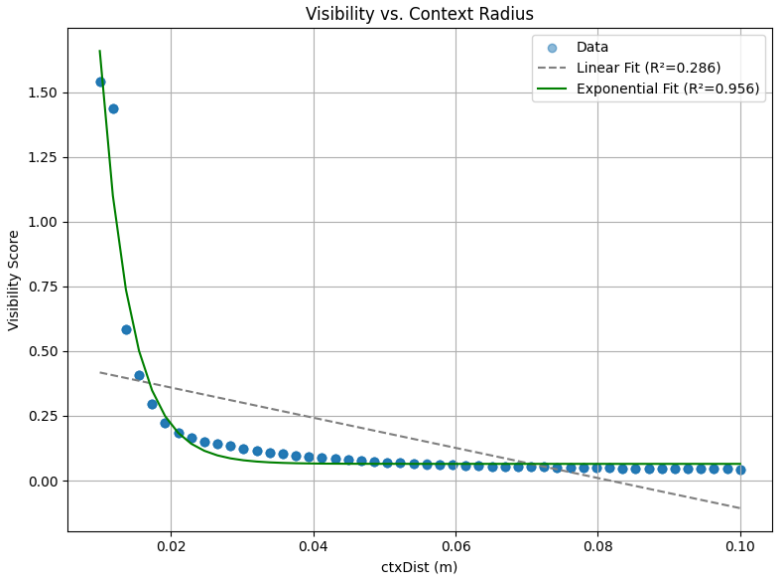


Figure 4.17: Model fits for visibility score vs. context radius (Stomach).

To illustrate the analysis pipeline, results for the stomach are shown in detail. Figure 4.17 compares the linear and exponential model fits to the visibility curve. The exponential curve clearly captures the initial steep decline and subsequent plateau in visibility scores. This organ also exhibited one of the largest LRT statistics, confirming a strongly nonlinear trend.

Figure 4.18 visualizes the result of knee detection using the Kneedle algorithm. The knee, located near $ctxDist = 0.041$ m, identifies the inflection point where visibility begins to plateau.

4.4.5 Statistical Analysis

A paired t-test between Kneedle and L-method values revealed a statistically significant difference ($p = 0.011$), with a mean deviation of 4.78 mm. This suggests that while both approaches produce comparable outputs, they do not select identical points on the visibility curve and may reflect slightly different interpretations of the curve's inflection.

A one-sample t-test comparing Kneedle estimates to the 2 cm baseline yielded $p = 0.350$, indicating no statistically significant deviation. This supports the hypoth-

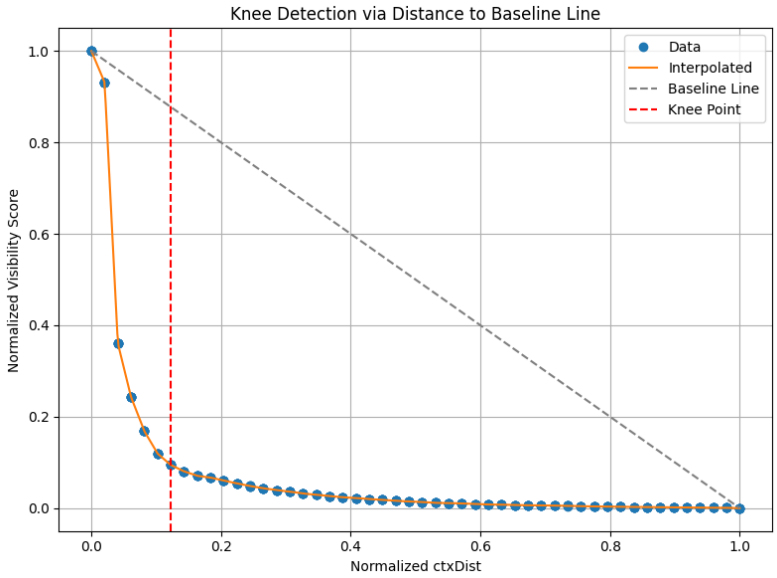


Figure 4.18: Knee point detection for the Stomach (Kneedle method).

esis that the Kneedle method identifies context distances aligned with perceptual expectations.

Table 4.7: Statistical comparison of context radius estimates.

Comparison	**Mean Difference (m)**	***p*-value**
Kneedle vs. L-method (paired)	0.00478	0.011
Kneedle vs. Manual (2 cm)	0.00426	0.350

These results validate the use of automatic knee detection for estimating optimal context radii in anatomical visualizations. The Kneedle algorithm, in particular, produced values statistically indistinguishable from the manually selected 2 cm baseline. Figure 4.19 visualizes the differences across methods for each organ, confirming the general agreement and trend alignment between them.

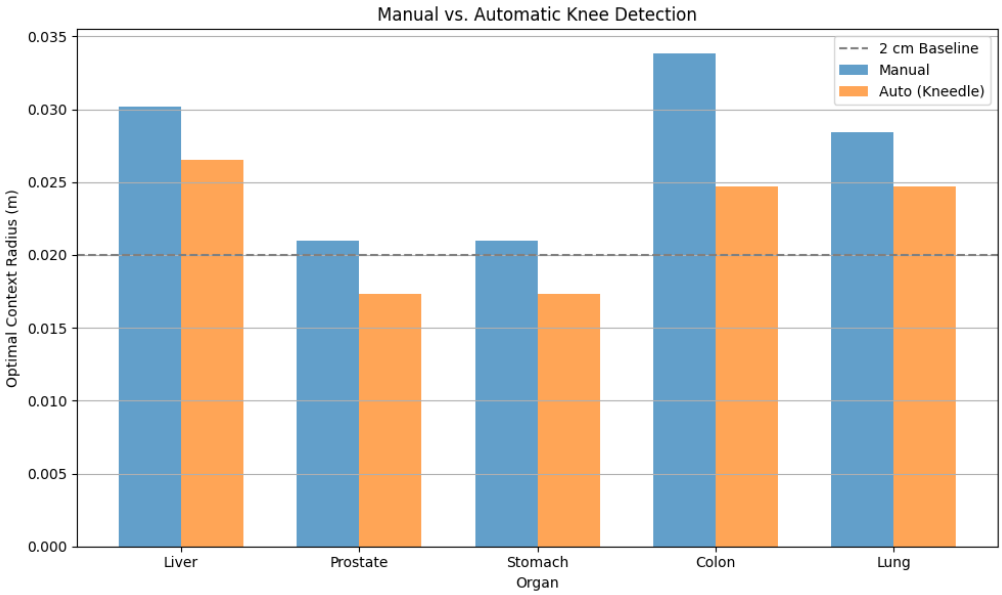


Figure 4.19: Comparison of context distance (in meters) selected by each method (Manual 2 cm baseline, Kneedle, L-method) across all evaluated organs

4.4.6 User interface

A user interface with an input text field as well as a button to trigger visualization is implemented as discussed in the Methodology chapter, the result can be seen in Figure 4.20. This implementation provides an all-in-one interaction for the system, a medical report can be typed or copied into the input field, and the button triggers the NER and visualization backend, without additional complications.

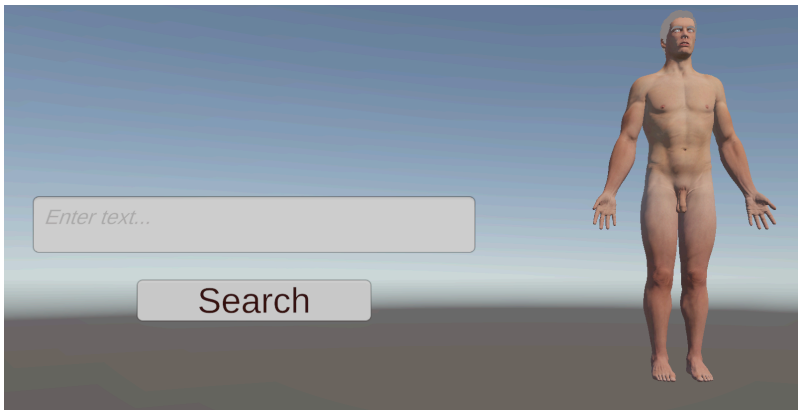


Figure 4.20: User Interface for Report Analysis

5

Conclusion

This thesis has explored the current NLP techniques available for capturing implicit anatomical entities. It was shown that simple rule-based systems fall short of extracting implicit concepts, and only recognize surface level mentions of biomedical entities. Transformers such as SapBERT were able to capture some deeper level context and have potential to provide good results with specific fine-tuning, but testing showed that available pretrained models do not provide satisfactory outcomes. LLM's such as LLaMA 2 can be run locally, and have capabilities of capturing implicit entities, but the novel UMLS-enhanced pipeline developed in this thesis provides the framework for controllable extraction of specific target entity types, based on relations between surface level disease mentions and the associated anatomical entities. Due to leveraging UMLS relations, it is capable of capturing nuanced relations in text that untrained baseline techniques lack. Visualization of these structures involved several parameters, and an algorithm for panning and zooming the camera based on the key structure's location and size was implemented. These techniques allowed a manual optimization of clarity and contextual relevance when representing anatomically relevant structures linked to pathological conditions. Further analysis was performed on the inclusion radius of surrounding structures, and it was shown that these experience diminishing returns for several organs, in a range that aligned with the manual optimization.

5.1 Discussion

Overall, the user interface, successful extraction of implicit entities through a novel implementation of a UMLS-backed entity recognition method, as well as thorough analysis and implementation of possible parameters to optimize visualization resulted in a functioning pipeline. The workflow included many aspects, such as the camera adjustment algorithm, to achieve a full system for visualization of clinical reports ready for clinical deployment to improve patient understanding of their conditions. The system presents an application of necessary enhancements of patient communication in the under explored field of visualizations in clinical settings, to improve compliance and thereby treatment outcomes. The research has shown that baseline NER techniques are limited in their capabilities of extracting implicit mentions, though LLM's show promise when fine tuning possibilities are considered. Surface level detection of entities, such as an integration of SciSpacy, falls short of adequately capturing the anatomical context in clinical reports, due to its implicit

nature. Transformer-based solutions, even when fine tuned on medical concepts, such as SapBERT-from-PubMedBERT-fulltext, produce unreliable and inaccurate results in the specific case explored in this thesis. The developed pipeline showed that the novel approach of leveraging UMLS relations provides an improved method of capturing implicit anatomical entities based on surface level disease mentions. The added advantage of the developed pipeline lies in its potential of selecting different NER labels and relations, therefore opening up the possibility of simple adaption to any use case of extracting implicit information in biomedical context. Visualization possibilities were thoroughly analyzed, and several adjustable parameters were implemented in a custom algorithm. Camera zooming and panning proved essential in optimizing visuals, as this facilitated a centralized and focused view on the target entity. Furthermore, surrounding structures proved to be the key to ensure clarity and contextual awareness in visual representations of clinical reports. A multitude of factors were considered, and a multi-parameter algorithm, with fine tuned parameters was implemented. A custom parameter, here referred to as context distance, was established, which takes into account the center to surface distance between surrounding objects, and provides an effective tool for establishing a balance between context and clutter. Analysis specifically on this parameter showed that there are diminishing returns for the inclusion radius of surrounding structures. In order to optimize the focus of visuals, parameters affecting the representation of included structures were also implemented and manually tuned. Control over opacity and dimming showed noticeable differences on the focus on the target entity after including surrounding structures. Thereby, this thesis has answered the research questions, by evaluating current NER methods in their capabilities of extracting implicit anatomical entities, proposing a novel pipeline with improved results, an analysis of the reliability of LLM based evaluation, an investigation and implementation of the key factors that influence a visualization's effectiveness, as well as an analysis on the diminishing returns of added surrounding structures.

Future work includes adding a female model, potential improvements to the pipeline, as well as later adaptions to other systems. Fuzzy string matching of UMLS concepts with model labels worked well, due to the semantic similarity between anatomical entities extracted and model labels. This is made possible because of the fine granularity of the model used, but this step may be improved with transformer based matching at the cost of added processing time. The entire UMLS release was setup in a local database, but many entities are unnecessary for individual use cases. Setting up the database with the subset SNOMED CT reduced the total size of the database, and therefore reduced processing time, yet does not include all anatomical entities of the full UMLS release.

Improvements of patient communication are a critical avenue of medical research, due to its potential of improving patient outcomes. Future prospects of patient communication include implementing NLP and AI techniques with robotics to create a medical assistant for patients. Developments such as Furhat Robotics' embodied conversational agents for face-to-face human agent interactions show promise in the possibilities of implementing a similar agent in a clinical setting [58]. Another field that shows promise in improving patient communication in clinical settings

are virtual reality based technologies. A 3D anatomical model is suitable for implementation in virtual reality, and companies such as Medicalholodeck have begun implementing educational apps for medical students, showcasing its potential to also be used in clinical settings for improved patient communication [59]. LLM's also showed promise in their capabilities of extracting implicit entities. An avenue of future work could focus on a locally installed fine-tuned LLM. While this involves computational resources and annotated data, better results than observed with LLaMA 2 may be achieved.

5.2 Conclusion

In conclusion, this thesis developed a novel UMLS-backed pipeline for implicit anatomical entity extraction that can be adapted to different use cases. A novel Unity logic with 4 adjustable parameters was implemented to optimize visual representations of anatomical entities, summarized in a user interface. Research showed that while existing NLP methods such as SciSpacy's en_core_sci_sm are effective tools for recognizing explicitly named entities, implicit entities remain hidden, and that leveraging structured relations provide more meaningful results. It was discovered that effective visualization techniques include panning and zooming the camera based on a highlighted structure's size and location, tuning a parameter for controlling which surrounding structures are included based on contextual distance, as well as tuning the level of dimming and opacity of these structures. These visualization parameters proved effective in enhancing visual representations by emphasizing relevant structures and minimizing visual clutter. Analysis on the inclusion radius showed that this parameter experiences diminishing returns at points that align with the manual selection. Previous work in the field of medical NLP has made evident that communication in clinical settings can be improved to increase patient compliance and improve treatment outcomes. Current efforts focus on medical text simplification, but personalized visualizations remain under explored, and show promise for providing the framework necessary for multimodal systems in clinical settings. Future prospects include conversational agents and virtual reality adaptions, but the necessary basis for these implementations requires further research in effective communication methods within clinical settings.

6

Ethics

This study employs large language models (LLMs), specifically LLaMA3 and Mistral, as automated evaluators of annotation quality. While this approach offers scalability and consistency, it also introduces ethical challenges. LLMs are trained on large, diverse, and often opaque datasets, which may encode biases, inaccuracies, or outdated clinical knowledge. Their judgments are not inherently transparent or verifiable, and their clinical reasoning may not align with domain expert standards. To mitigate these concerns, a validation experiment was conducted to test whether the LLMs could reliably distinguish between meaningful (gold-standard) and nonsensical annotations. High discrimination performance (AUC > 0.91) and large effect sizes were observed, supporting their use as evaluators in this context. Nevertheless, the use of LLMs as proxies for expert judgment must be treated with caution. They are not substitutes for clinical reviewers, and future deployments of this evaluation method should involve regular human auditing and benchmarking against expert consensus.

A significant ethical limitation of this work lies in the exclusive use of a male anatomical model when mapping extracted findings to anatomical sites. This choice was based on tool availability and alignment with existing SNOMED CT mapping structures. However, it risks perpetuating a systemic bias that overlooks anatomical variation across sexes and gender identities. Relying solely on male anatomy in clinical NLP pipelines can lead to model blind spots. Particularly in areas like reproductive health, pelvic anatomy, and differential disease presentation the considerations for female anatomy are critical. Such bias can inadvertently reinforce healthcare disparities, especially for female, intersex, and transgender patients whose anatomy may deviate from the assumed norm. This limitation is acknowledged as a significant area for future work. Expanding anatomical linkage frameworks to incorporate sex-specific and inclusive anatomical representations is essential to ensure fairness, generalizability, and ethical clinical deployment. Any future clinical application of this pipeline must explicitly address these representational gaps before deployment in real-world settings.

6.1 Licenses and Resources

The following models, libraries, and datasets were used during the course of this research. Each is credited below with the appropriate license and source where

applicable.

6.1.1 Natural Language Processing Models

- **SciSpaCy** `en_core_sci_sm` and `en_ner_bc5cdr_md`
  Biomedical NLP models from SciSpaCy, distributed under the MIT License.
  `https://allenai.github.io/scispacy`

- **SapBERT from PubMedBERT (fulltext)**
  Biomedical concept encoder trained on UMLS terms. Released under the Apache 2.0 License.
  `https://github.com/cambridgeltl/sapbert`

6.1.2 Large Language Models

- **LLaMA 2 and LLaMA 3**
  Foundation models developed by Meta AI. Research use subject to the LLaMA research license agreement.
  `https://ai.meta.com/llama/`

- **Mistral**
  Open-weight language model released by Mistral AI. Licensed under Apache 2.0.
  `https://mistral.ai/news/`

6.1.3 Visualization and Software Frameworks

- **Unity Engine**
  Used for 3D anatomical visualization and rendering. Licensed under Unity Personal/Pro depending on user agreement.
  `https://unity.com`

6.1.4 Clinical Ontologies

- **UMLS (Unified Medical Language System)**
  Utilized for medical concept normalization. Accessed under the UMLS Metathesaurus License from the U.S. National Library of Medicine.
  `https://www.nlm.nih.gov/research/umls`

- **SNOMED CT**
  Clinical terminology used for medical concept mapping. Distributed under SNOMED International licensing; use subject to country-specific license agreements.
  `https://www.snomed.org/snomed-ct`

6.1.5 Python Libraries

The following Python packages were used in the implementation of data processing, statistical analysis, and figure generation:

- **NumPy** [60] numerical operations and array manipulation
- **Pandas** tabular data handling and CSV I/O
- **Matplotlib** and **Seaborn** data visualization and plotting
- **SciPy** statistical testing, curve fitting, and numerical optimization
- **scikit-learn** regression models and R^2 metric computation

Bibliography

[1] M. R. Andrus and M. T. Roth, "Health literacy: A review," *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy*, vol. 22, no. 3, pp. 282–302, 2002. DOI: https://doi.org/10.1592/phco.22.5.282.33191. eprint: https://accpjournals.onlinelibrary.wiley.com/doi/pdf/10.1592/phco.22.5.282.33191. [Online]. Available: https://accpjournals.onlinelibrary.wiley.com/doi/abs/10.1592/phco.22.5.282.33191.

[2] J. Lossio-Ventura, S. Boussard, J. Morzan, and T. Hernandez-Boussard, "Clinical named-entity recognition: A short comparison," in *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Epub 2020 Feb 6, Nov. 2019, pp. 1548–1550. DOI: 10.1109/BIBM47256.2019.8983406.

[3] G. Brajnik, "A comparative test of web accessibility evaluation methods," in *Proceedings of the 10th international ACM SIGACCESS conference on Computers and accessibility - Assets '08*, New York, New York, USA: ACM Press, 2008, p. 113, ISBN: 9781595939760. DOI: 10.1145/1414471.1414494. [Online]. Available: http://portal.acm.org/citation.cfm?doid=1414471.1414494.

[4] A. Kantor. "The health literacy advisor." (), [Online]. Available: https://www.healthliteracyinnovations.com/products/hla. (accessed: 14.12.2024).

[5] V. Body, *Visible body suite*, Interactive 3D anatomy visualization tool. Accessed: 2025-01-13, 2025. [Online]. Available: https://www.visiblebody.com/.

[6] M. Lewis, Y. Liu, N. Goyal, *et al.*, *Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension*, 2019. arXiv: 1910.13461 [cs.CL]. [Online]. Available: https://arxiv.org/abs/1910.13461.

[7] L. Martin, A. Fan, É. de la Clergerie, A. Bordes, and B. Sagot, *Muss: Multilingual unsupervised sentence simplification by mining paraphrases*, 2021. arXiv: 2005.00352 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2005.00352.

[8] A. Devaraj, I. Marshall, B. Wallace, and J. Li, "Paragraph-level simplification of medical texts," vol. 2021, Jun. 2021, pp. 4972–4984. DOI: 10.18653/v1/2021.naacl-main.395.

[9] C. M. Ardila, D. González-Arroyave, and M. Zuluaga-Gómez, "Efficacy of three-dimensional models for medical education: A systematic scoping review of randomized clinical trials," *Heliyon*, vol. 9, no. 2, e13395, 2023. DOI: 10.1016/j.heliyon.2023.e13395.

[10] Y. Xu, J. Hua, Z. Ni, *et al.*, "Anatomical entity recognition with a hierarchical framework augmented by external resources," *PLoS ONE*, vol. 9, no. 10, e108396, Oct. 2014. DOI: 10.1371/journal.pone.0108396.

[11] K. Pakhale, *Comprehensive overview of named entity recognition: Models, domain-specific applications and challenges*, 2023. arXiv: 2309.14084 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2309.14084.

[12] I. Keraghel, S. Morbieu, and M. Nadif, *Recent advances in named entity recognition: A comprehensive survey and comparative study*, 2024. arXiv: 2401.10825 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2401.10825.

[13] M. Neumann, D. King, I. Beltagy, and W. Ammar, *Scispacy: Fast and robust models for biomedical natural language processing*, https://allenai.github.io/scispacy/, Accessed: 2025-04-21, 2019.

[14] F. Liu, E. Shareghi, Z. Meng, M. Basaldella, and N. Collier, "Self-alignment pretraining for biomedical entity representations," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Online: Association for Computational Linguistics, Jun. 2021, pp. 4228–4238. [Online]. Available: https://www.aclweb.org/anthology/2021.naacl-main.334.

[15] H. Touvron, L. Martin, K. Stone, *et al.*, *Llama 2: Open foundation and fine-tuned chat models*, 2023. arXiv: 2307.09288 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2307.09288.

[16] R. Leaman, R. Khare, and Z. Lu, "Challenges in clinical natural language processing for automated disorder normalization," *Journal of Biomedical Informatics*, vol. 57, pp. 28–37, Oct. 2015, Epub 2015 Jul 14. DOI: 10.1016/j.jbi.2015.07.010.

[17] O. Bodenreider, "The unified medical language system (umls): Integrating biomedical terminology," *Nucleic Acids Research*, vol. 32, no. Database issue, pp. D267–D270, 2004. DOI: 10.1093/nar/gkh061. [Online]. Available: https://doi.org/10.1093/nar/gkh061.

[18] D. Fraile Navarro, K. Ijaz, D. Rezazadegan, *et al.*, "Clinical named entity recognition and relation extraction using natural language processing of medical free text: A systematic review," *International Journal of Medical Informatics*, vol. 177, p. 105 122, 2023, ISSN: 1386-5056. DOI: https://doi.org/10.1016/j.ijmedinf.2023.105122. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1386505623001405.

[19] A. Jolly, V. Pandey, I. Singh, and N. Sharma, "Exploring biomedical named entity recognition via scispacy and biobert models," *Open Biomedical Engineering Journal*, vol. 18, e18741207289680, 2024. DOI: 10.2174/0118741207289680240510045617. [Online]. Available: http://dx.doi.org/10.2174/0118741207289680240510045617.

[20] J. Lee, W. Yoon, S. Kim, *et al.*, "Biobert: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020. DOI: 10.1093/bioinformatics/btz682. [Online]. Available: https://academic.oup.com/bioinformatics/article/36/4/1234/5566506.

[21] S. Tsang, *Brief review: Biobert a pre-trained biomedical language representation model for biomedical text mining*, https://sh-tsang.medium.com/

brief-review-biobert-a-pre-trained-biomedical-language-representation-model-for-biomedical-text-4b5cf07efdd7, Accessed: 2025-04-21, 2020.

[22] Q. Chen, Y. Hu, X. Peng, *et al.*, "Benchmarking large language models for biomedical natural language processing applications and recommendations," *Nature Communications*, vol. 16, p. 3280, 2025. DOI: 10.1038/s41467-025-56989-2. [Online]. Available: https://doi.org/10.1038/s41467-025-56989-2.

[23] Y. Hu, Q. Chen, J. Du, *et al.*, *Improving large language models for clinical named entity recognition via prompt engineering*, 2024. arXiv: 2303.16416 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2303.16416.

[24] Q. Chen, Y. Hu, X. Peng, *et al.*, "Benchmarking large language models for biomedical natural language processing applications and recommendations," *Nature Communications*, vol. 16, p. 3280, 2025. DOI: 10.1038/s41467-025-56989-2. [Online]. Available: https://doi.org/10.1038/s41467-025-56989-2.

[25] Y. Hu, X. Zuo, Y. Zhou, *et al.*, *Information extraction from clinical notes: Are we ready to switch to large language models?* 2025. arXiv: 2411.10020 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2411.10020.

[26] S. A. R. team, *Sfrjudge: Judge models built with meta llamaă3 and mistral nemo*, Blog post, Presentation of singlepoint, pairwise, and Likertscale judge models built on Llamaă3 / Mistral NeMO architecture, 2024.

[27] A. S. Thakur, K. Choudhary, V. S. Ramayapally, S. Vaidyanathan, and D. Hupkes, "Judging the judges: Evaluating alignment and vulnerabilities in llms-asjudges," in *ICLR Workshops (Conference Submission)*, Study of multiple LLM judge models, biases and alignment to human scores, 2024.

[28] Encord, *Llm as a judge: What it is, how it works, and limitations*, Blog post, Describes how structured and chain-of-thought prompts improve LLM evaluation reliability, Apr. 2024. [Online]. Available: https://encord.com/blog/llm-as-a-judge/.

[29] SuperAnnotate, *Llm evaluation: Metrics, frameworks, and best practices*, Blog post, Comprehensive evaluation guide covering metrics, frameworks, tools, and best practices for LLMs, Jun. 2025.

[30] Snorkel AI Team, *Llmasajudge for enterprises: Evaluate model alignment at scale*, Blog post, Discusses prompt design, SME-in-the-loop validation, and best practices for LLM-as-a-judge systems, 2025. [Online]. Available: https://snorkel.ai/llm-as-judge-for-enterprises/.

[31] U.S. National Library of Medicine, *Umls semantic network*, Accessed: 2025-05-25, n.d. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK9679/.

[32] K. Stanceski, S. Zhong, X. Zhang, *et al.*, "The quality and safety of using generative ai to produce patient-centred discharge instructions," *npj Digital Medicine*, vol. 7, p. 329, 2024. DOI: 10.1038/s41746-024-01336-w. [Online]. Available: https://doi.org/10.1038/s41746-024-01336-w.

[33] A. F. Glick, C. Brach, H. S. Yin, and B. P. Dreyer, "Health literacy in the inpatient setting: Implications for patient care and patient safety," *Pediatric*

Clinics of North America, vol. 66, no. 4, pp. 805–826, Aug. 2019, Epub 2019 May 23. DOI: 10.1016/j.pcl.2019.03.007.

[34] K. Jubbal, S. Chun, J. Chang, S. Zhang, L. Terrones, and J. S. Huang, "Parental and youth understanding of the informed consent process for pediatric endoscopy," *Journal of Pediatric Gastroenterology and Nutrition*, vol. 60, no. 6, pp. 769–775, Jun. 2015. DOI: 10.1097/MPG.0000000000000719.

[35] D. Vernon, J. E. Brown, E. Griffiths, A. M. Nevill, and M. Pinkney, "Reducing readmission rates through a discharge follow-up service," *Future Healthcare Journal*, vol. 6, no. 2, pp. 114–117, Jun. 2019. DOI: 10.7861/futurehosp.6-2-114.

[36] J. P. Lalor, W. Hu, M. Tran, H. Wu, K. M. Mazor, and H. Yu, "Evaluating the effectiveness of noteaid in a community hospital setting: Randomized trial of electronic health record note comprehension interventions with patients," *Journal of Medical Internet Research*, vol. 23, no. 5, e26354, May 2021. DOI: 10.2196/26354.

[37] B. Ondov, K. Attal, and D. Demner-Fushman, "A survey of automated methods for biomedical text simplification," *Journal of the American Medical Informatics Association*, vol. 29, no. 11, pp. 1976–1988, Oct. 2022. DOI: 10.1093/jamia/ocac149. [Online]. Available: https://doi.org/10.1093/jamia/ocac149.

[38] S. Kandula, D. Curtis, and Q. Zeng-Treitler, "A semantic and syntactic text simplification tool for health content," in *AMIA Annual Symposium Proceedings*, American Medical Informatics Association, 2010, pp. 366–370. [Online]. Available: https://www.researchgate.net/publication/49967676_A_Semantic_and_Syntactic_Text_Simplification_Tool_for_Health_Content.

[39] K. Jeblick, B. Schachtner, J. Dexl, *et al.*, *Chatgpt makes medicine easy to swallow: An exploratory case study on simplified radiology reports*, 2022. arXiv: 2212.14882 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2212.14882.

[40] C. DeSai, K. Janowiak, B. Secheli, *et al.*, "Empowering patients: Simplifying discharge instructions," *BMJ Open Quality*, vol. 10, no. 3, e001419, Sep. 2021. DOI: 10.1136/bmjoq-2021-001419. [Online]. Available: https://doi.org/10.1136/bmjoq-2021-001419.

[41] A. Mishra, P. Russell, and E. Niebur, "A proto-object based saliency model in three-dimensional space," *Journal of Vision*, vol. 14, no. 12, p. 3, 2014, Demonstrates that incorporating depth cues in a proto-object saliency model improves alignment with 3D human fixation data.

[42] H. Rahimi-Nasrabadi, V. Moore-Stoll, J. Tan, *et al.*, "Luminance contrast shifts dominance balance between on and off pathways in human vision," *Journal of Neuroscience*, vol. 43, no. 6, pp. 993–1007, 2023. DOI: 10.1523/JNEUROSCI.1672-22.2022. [Online]. Available: https://doi.org/10.1523/JNEUROSCI.1672-22.2022.

[43] J. M. Wolfe, "Guided search 2.0: A revised model of visual search," *Psychonomic Bulletin & Review*, vol. 1, no. 2, pp. 202–238, 1994.

[44] R. Rosenholtz, Y. Li, and L. Nakano, "Measuring visual clutter," *Journal of Vision*, vol. 7, no. 2, pp. 17–17, 2007.

[45] E. Peli, "Contrast in complex images," *Journal of the Optical Society of America A*, vol. 7, no. 10, pp. 2032–2040, 1990.

[46] A. E. W. Johnson, L. Bulgarelli, T. Pollard, S. Horng, L. A. Celi, and R. G. Mark, *Mimiciv (version 2.2)*, PhysioNet, RRID:SCR_007345, 2023. [Online]. Available: https://doi.org/10.13026/6mm1%E2%80%91ek67.

[47] F. Liu, E. Shareghi, Z. Meng, M. Basaldella, and N. Collier, "Self-alignment pretraining for biomedical entity representations," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, K. Toutanova, A. Rumshisky, L. Zettlemoyer, *et al.*, Eds., Online: Association for Computational Linguistics, Jun. 2021, pp. 4228–4238. DOI: 10.18653/v1/2021.naacl-main.334. [Online]. Available: https://aclanthology.org/2021.naacl-main.334/.

[48] M. Neumann, D. King, I. Beltagy, and W. Ammar, "ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing," in *Proceedings of the 18th BioNLP Workshop and Shared Task*, Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 319–327. DOI: 10.18653/v1/W19-5034. eprint: arXiv:1902.07669. [Online]. Available: https://www.aclweb.org/anthology/W19-5034.

[49] National Library of Medicine (US), *UMLS Knowledge Sources [dataset on the Internet]*, http://www.nlm.nih.gov/research/umls/licensedcontent/umlsknowledgesources.html, Release 2024AA. Bethesda (MD): National Library of Medicine (US); 2024 May 6 [cited 2024 Jul 15], 2024.

[50] J. Gu, X. Jiang, Z. Shi, *et al.*, *A survey on llm-as-a-judge*, 2025. arXiv: 2411.15594 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2411.15594.

[51] Meta AI, *Llama 3: Open foundation and instruction-tuned models*, https://ai.meta.com/llama/, Accessed: July 2025, 2024.

[52] M. AI, *Mistral 7b*, https://mistral.ai/news/announcing-mistral-7b/, Accessed: July 2025, 2023.

[53] International Telecommunication Union, *Parameter values for the HDTV standards for production and international programme exchange*, ITU-R Recommendation BT.709-6, https://www.itu.int/rec/R-REC-BT.709, 2015.

[54] T. Drew, M. L. Vo, and J. M. Wolfe, "The invisible gorilla strikes again: Sustained inattentional blindness in expert observers," *Psychological Science*, vol. 24, no. 9, pp. 1848–1853, 2013.

[55] M. S. Tsai, J. H. Siewerdsen, J. W. Stayman, E. Asma, and D. J. Tward, "Tradeoffs between context and visibility in volumetric image display: Implications for user interface design in imageguided procedures," *Medical Physics*, vol. 47, no. 7, pp. 3033–3046, 2020. DOI: 10.1002/mp.14191.

[56] E. Coucke, H. Akbar, A. Kahloon, *et al.*, "Biliary obstruction," in *StatPearls [Internet]*, Updated 2022 Nov 26, Treasure Island (FL): StatPearls Publishing, 2025. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK539698/.

[57] L. Grant and S. John, "Cholestatic jaundice," in *StatPearls*. Treasure Island (FL): StatPearls Publishing, 2025, [Updated 2025 Jan 19].

[58] S. Al Moubayed, J. Beskow, G. Skantze, and B. Granström, "Furhat: A back-projected human-like robot head for multiparty human-machine interaction," in *Cognitive Behavioural Systems*, A. Esposito, A. M. Esposito, A. Vinciarelli, R. Hoffmann, and V. C. Müller, Eds., ser. Lecture Notes in Computer Science, vol. 7403, Springer, Berlin, Heidelberg, 2012, pp. 114–130. DOI: 10.1007/978-3-642-34584-5_9.

[59] J. Arensmeyer, B. Bedetti, P. Schnorr, *et al.*, "A system for mixed-reality holographic overlays of real-time rendered 3d-reconstructed imaging using a video pass-through head-mounted displaya pathway to future navigation in chest wall surgery," *Journal of Clinical Medicine*, vol. 13, no. 7, p. 2080, 2024. DOI: 10.3390/jcm13072080. [Online]. Available: https://doi.org/10.3390/jcm13072080.

[60] C. R. Harris, K. J. Millman, S. J. van der Walt, *et al.*, "Array programming with NumPy," *Nature*, vol. 585, no. 7825, pp. 357–362, Sep. 2020. DOI: 10.1038/s41586-020-2649-2. [Online]. Available: https://doi.org/10.1038/s41586-020-2649-2.