



SCHOOL OF COMPUTATION, INFORMATION
AND TECHNOLOGY - INFORMATICS

TECHNICAL UNIVERSITY OF MUNICH

Bachelor's Thesis in Information Systems

**Evaluating and Enhancing Location-Aware
Visual Document Segmentation for Oncology
Guidelines**

Matteo Felipe Merz



SCHOOL OF COMPUTATION, INFORMATION
AND TECHNOLOGY - INFORMATICS

TECHNICAL UNIVERSITY OF MUNICH

Bachelor's Thesis in Information Systems

Evaluating and Enhancing Location-Aware Visual Document Segmentation for Oncology Guidelines

Evaluierung und Erweiterung positioneller visueller Dokumentensegmentierung für Onkologie-Richtlinien

Author:	Matteo Felipe Merz
Supervisor:	Prof. Dr. Florian Matthes
Advisor:	Jonas Gottal, M.Sc.; Juraj Vladika, M.Sc.
Submission Date:	22.02.2026

I confirm that this bachelor's thesis in information systems is my own work and I have documented all sources and material used.

Location, Submission Date

Author

AI Assistant Usage Disclosure

Introduction

Performing work or conducting research at the Chair of Software Engineering for Business Information Systems (sebis) at TUM often entails dynamic and multi-faceted tasks. At sebis, we promote the responsible use of *AI Assistants* in the effective and efficient completion of such work. However, in the spirit of ethical and transparent research, we require all student researchers working with sebis to disclose their usage of such assistants.

For examples of correct and incorrect AI Assistant usage, please refer to the original, unabridged version of this form, located at [this link](#).

Use of *AI Assistants* for Research Purposes

I have used AI Assistant(s) for the purposes of my research as part of this thesis.

Yes No

Explanation:

I confirm in signing below, that I have reported all usage of AI Assistants for my research, and that the report is truthful and complete.

Location, Date

Author

Acknowledgments

Abstract

Kurzfassung

Contents

Acknowledgments	iv
Abstract	v
Kurzfassung	vi
1. Introduction	1
1.1. Problem Statement	1
1.2. Objectives	2
2. Foundations	4
2.1. Oncology guideline documents	4
2.2. Natural Language Processing Fundamentals	5
2.2.1. Tokenization	5
2.2.2. Sentence Embeddings	6
2.2.3. Cosine Similarity	6
2.3. Vision-Language Models	7
2.3.1. Bounding Boxes	7
2.3.2. Intersection over Union	7
2.4. Retrieval-Augmented Generation	8
2.4.1. Architecture of RAG Systems	8
2.4.2. Indexing	9
2.4.3. Source Attribution	10
2.5. Document Parsing	10
2.5.1. Modular Pipeline Systems	11
2.5.2. End-to-End VLM models	12
2.6. Chunking	13
2.6.1. Window Passages	13
2.6.2. Semantic Passages	14
2.6.3. Discourse Passages	14
2.6.4. Metadata Attachments	14
3. Methodology	15
3.1. Pipeline Overview	15
3.2. Data Representations	16
3.2.1. ParsingResultType	16
3.2.2. ParsingBoundingBox	16

3.2.3.	ParsingResult	16
3.2.4.	ChunkingResult	17
3.2.5.	Chunk	18
3.3.	Parsing Module	18
3.3.1.	Unstructured.io	20
3.3.2.	Docling	21
3.3.3.	MinerU	21
3.3.4.	Gemini 2.5 Flash	22
3.3.5.	LlamaParse	22
3.3.6.	Google Document AI LayoutParser	23
3.4.	Chunking Module	23
3.4.1.	Fixed-Size Chunking	24
3.4.2.	Recursive Character Chunking	25
3.4.3.	Breakpoint-based Semantic Chunking	25
3.4.4.	Hierarchical Chunking	26
3.5.	Evaluation Framework	26
3.5.1.	Document Layout Analysis Evaluation	26
3.5.2.	Content Extraction Evaluation	28
3.5.3.	Chunking Evaluation	32
3.5.4.	Evaluation Environment	34
4.	Results	35
5.	Discussion	38
6.	Conclusion	39
A.	General Addenda	40
	List of Figures	44
	List of Tables	45
	Acronyms	46

1. Introduction

1.1. Problem Statement

Clinical practice guidelines (CPGs) are fundamental to the efficient and reliable treatment of various illnesses [cpg_trust]. Oncology guidelines are a subgroup of these documents, revolving around the treatment of various forms of cancer [nccn_about_guidelines]. CPGs not only aid doctors in deciding on the optimal treatment options but also support patients in understanding their illness. In recent years, due to advancements in technology, novel therapeutics, and personalized medicine, clinical guidelines have drastically increased in size and complexity [guidelines_length]. As of 2019, the average oncology guideline published by the National Comprehensive Cancer Network (NCCN) was 198 pages long, showing an annual increase of 7.5 percent over the previous 23 years [guidelines_length]. This increase of complexity forces medical personnel to invest more time in order to be able to provide optimal care for cancer patients. Especially for individual practitioners this additional strain might become unsustainable if complexity continues to increase [guidelines_length].

The Aidvice project proposes to address this problem by leveraging recent advantages in artificial intelligence (AI). Specifically, the project revolves around the development of a retrieval-augmented generation (RAG) based knowledge assistant [aidvice]. RAG is an emerging paradigm which addresses a fundamental problem of traditional large language models (LLMs) [rag]. While LLMs excel at many natural language processing (NLP) tasks, they are prone to “hallucinations” and inaccurate answers, when sought information goes beyond the model’s training data [rag_survey]. This provides a major obstacle for the usage of LLMs in the medical field, where accurate and reliable answers are of the highest priority [lins]. RAG mitigates these drawbacks by retrieving additional context from an external knowledge source which the LLM can take advantage of during answer generation [chunk_size_effect_on_rag, rag].

The efficiency of such a RAG system is fundamentally constrained by the quality and relevance of the context retrieved from the knowledge base [dense_x, data_quality_rag]. Therefore, the construction of the knowledge base out of the oncology guidelines is a critical aspect of the project. Additionally, as the project has a clear focus on verifiability and traceability, there is an additional requirement to provide visual source attribution with the models responses. This means that retrieved passages need to include accurate positional information, giving visual confirmation to the practitioner about the origin of the retrieved context [visa].

As LLMs are constrained by the size of their context window, it is not feasible to store the entire guideline documents as individual entries in the knowledge base [rag_survey]. Therefore, the documents need to be split up into smaller text chunks that fit into the model’s

context window [**rag_survey**]. This process is called chunking [**semantic_chunking**].

The guideline documents are stored in the unstructured portable document format (PDF). In order to be further processed for the knowledge base, they first need to be transformed into a machine-readable structured data format through a process called document parsing (DP) [**parsingunveiled**]. The inherent structure of the guidelines poses multiple challenges for this process, such as complex tables, varying layouts and occasional formatting errors.

During retrieval the model identifies the most relevant passages in the knowledge base based on their similarity to the user's query [**rag**]. If the stored text chunks are too long, important information might be lost between irrelevant details [**dense_x, data_quality_rag**]. On the other hand, storing too short text chunks can result in important statements being broken up into multiple chunks and losing their meaning. In order to maximize the quality of the retrieved chunks, both the chunk size as well as the chunking strategy, used to decide where to split up the oncology guidelines, need to be optimized [**data_quality_rag**].

Additionally, established implementations of popular chunking strategies do not fulfill the requirement of visual source attribution at the granularity required by the Aidvice project [**langchain, llamaindex, docling_toolkit, langchain_splitting_recursively**]. Therefore there is a need for the development of a novel solution that addresses this issue.

1.2. Objectives

This study addresses three fundamental research questions regarding the data preparation for a RAG based knowledge assistant for oncology guidelines. Each of the following research questions addresses a specific aspect of the evaluation and improvement of the document segmentation process required for the construction of the knowledge base.

- **RQ1:** How are the challenges introduced by oncology guidelines reflected in established benchmarks for document parsing?
- **RQ2:** Which metrics are most useful to measure the effectiveness of document parsing and chunking methods?
- **RQ3:** How can current segmentation methods be adapted or expanded on to fulfill the requirements of a RAG based knowledge assistant with visual source attribution?

To identify the challenges posed by the oncology guidelines, we perform a qualitative analysis identifying the characteristics of oncology guidelines that are relevant to the DP process, such as their formatting, layout and common types of structural elements. We then identify established document benchmarks and datasets which contain documents that most closely resemble these characteristics. Through this analysis, we underline the transferability of results achieved on these datasets to our application, while identifying unrepresented characteristics which require manual comparisons. This approach allows the evaluation of various DP techniques on established benchmarks, without the availability of a dedicated oncology guideline benchmark.

In order to evaluate the effectiveness of both document parsing and chunking strategies, we identify various metrics used in existing literature. We then perform a comparative analysis of the identified metrics, evaluating their suitability for our application. Based on this analysis, we select a set of metrics which are most suitable for our evaluation.

Finally, we propose a novel solution for the visual source attribution requirement of the Aidvice project that enables the traceability of the chunks content to its constituent text on a higher granularity than existing chunking implementations. We adapt and expand on existing chunking strategies in order to provide accurate positional information for each text chunk. By introducing a universal data format for the output of the DP implementations, we enable the direct comparison of various document parsing techniques using the benchmarks and metrics identified in RQ1 and RQ2. Through this evaluation we identify promising combinations of document parsing and chunking strategies for the creation of the knowledge base of the Aidvice project, while providing a modular framework for future experiments and improvements.

2. Foundations

2.1. Oncology guideline documents

CPGs help improve patient care by giving recommendations on the optimal treatment and prevention of various diseases [**cpg_good_bad_ugly**, **oncology_review**]. They are developed by groups of independent multi-disciplinary experts and are based on a robust systematic review of available treatment options and knowledge gained from clinical experience [**cpg_trust**, **cpg_good_bad_ugly**, **oncology_review**]. Instead of dictating a single definitive treatment option, CPGs instead focus on aiding the decision making process, promoting treatment options with proven benefits and discouraging ineffective or harmful treatments [**cpg_trust**, **cpg_good_bad_ugly**]. As such, they aim to improve the quality of the provided health care by encouraging the translation of research into medical practice [**oncology_review**]. Oncology guidelines are a subgroup of this document type, focusing on the treatment and rehabilitation options for various types of cancers [**oncology_review**].

In order to further improve the quality and standardization of oncology guidelines [**oncology_review**, **guidelines_length**], and therefore cancer care, several prominent organizations have emerged, which endorse and publish selected oncology guidelines. Prominent examples include the NCCN [**nccn_about_guidelines**], the European Society for Medical Oncology (ESMO) [**esmo**], and, for oncology guidelines in the German language, the “Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften” (AWMF) [**awmf**]. Over the last decades oncology has seen many advances in the research and treatment outcomes of many forms of cancers [**guidelines_length**, **advancements_oncology**]. Following these findings and advancements, the number of available treatment options has increased drastically [**guidelines_length**]. This increase in available treatments is ultimately reflected in the increasing complexity of oncology guidelines. **guidelines_length** found that, between 1996 and 2019, the mean page count of guidelines published by the NCCN has increased from 26 to 198 pages, with the number of referenced citations per guideline also increasing from an average of 30 to 111.

In order to identify common characteristics between the layout, typography and page design of different oncology guideline documents, we perform a qualitative analysis on a selection of german and english guideline documents from multiple publishing organizations. Despite significant variability between guidelines from different publishers, several shared characteristics can be observed:

Data format: The primary data format for digital distribution of oncology guidelines is the PDF. PDF is a data format designed to enable the reliable distribution and viewing

of electronic documents independent of the viewing or creating environment [pdf_iso]. Particularly, these documents are born-digital PDF files, created through digital processes, instead of scanning analog documents.

Page geometry: All observed documents are provided in the standard A4 format, thereby sharing common page dimensions. While the majority of oncology guidelines are provided in a vertical orientation, both horizontal and mixed page orientations are possible. Additionally, there exist some cases where two neighboring vertical pages are contained in a single horizontal page.

Content and layout: The formatting and content of the guideline documents is heavily dependent on their target audience. “Standard” guideline documents, addressing medical professionals, resemble typical scientific documents. They are mostly provided in a single or double-column layout, and, due to their focus on aggregating the results of previous studies, predominantly text-heavy. Additionally, they often contain complex tables which may span multiple pages, primarily to compare different treatment options against each other. While less frequent, figures, mathematical formulas and images are also occasionally included. As CPGs are often too complicated for patients to understand, some publishers provide “patient guidelines” alongside their CPGs. These documents translate the recommendations from the CPG into a language that is understood by the general population, while leaving out scientific details that are less relevant to the patient. Compared to the CPG these documents usually incorporate more figures and visual elements while offering more variability in their typography and page designs.

Document quality: Depending on the CPG’s age and publishing organization, the formatting of the document may contain significant structural errors. Observed formatting issues include overlapping text, empty pages between content, tables extending into the page margins, and invisible text on document pages.

2.2. Natural Language Processing Fundamentals

According to **nlp_advances**, NLP “employs computational techniques for the purpose of learning, understanding, and producing human language content” (p.1). The introduction of the transformer architecture by **transformer** and the subsequent development of LLMs has revolutionized the field in recent years [llm_in_nlp]. In order to understand how LLMs process and perceive information, it is necessary to examine various fundamental concepts.

2.2.1. Tokenization

Tokenization refers to the segmentation of text into sub-word units called tokens [fast_tokenization]. Tokens are the fundamental text representation for most NLP tasks. With a granularity located between characters and words, tokens can retain linguistic meaning while also being able

to represent arbitrary text with a relatively concise vocabulary [**fast_tokenization**]. Using tokenization any given text can essentially be represented as a list of integers, with each integer being the identifier to a specific token in the tokenizer's dictionary [**token_history**]. During training, the tokenizer creates its dictionary by finding character pairings that occur with the highest frequency in the training data [**token_history**]. Additionally, with the multitude of different techniques for modern sub-word tokenization [**wordpiece**, **sentencepiece**, **unigram**], the same input text can lead to drastically different outputs depending on the specific tokenizer and training data. Therefore, tokenizers always need to match the NLP models they are used with.

2.2.2. Sentence Embeddings

Sentence embeddings encode the semantical meaning of sentences into vectors of fixed-dimensionality [**speech_and_language**]. Every modern NLP algorithm uses embeddings as the representation of the meaning of texts [**speech_and_language**]. Using this technique, the meaning of the text is transformed into a machine-understandable format, with the embedding vectors of closely related sentences being closer to each other in the vector space.

2.2.3. Cosine Similarity

The cosine similarity determines the similarity between two sentences by calculating the cosine of the angle between the embedding vectors v and w . Cosine similarity builds on top of the dot product metric (Equation 2.1). The dot product tends to be high when v and w have large values in the same dimensions, therefore measuring their similarity [**speech_and_language**].

$$\text{dot product}(v, w) = v \cdot w = \sum_{i=1}^N v_i w_i = v_1 w_1 + v_2 w_2 + \dots + v_N w_N \quad (2.1)$$

However, the dot product is not invariant to the length of the vector, defined in Equation 2.2, producing higher values for vectors of greater length [**speech_and_language**]. This leads to skewed similarity values if vectors are not normalized prior.

$$|v| = \sqrt{\sum_{i=1}^N v_i^2} \quad (2.2)$$

The cosine similarity is calculated as the normalized dot product, as defined in Equation 2.3. As such, it is a measurement of the similarity between two vectors that is invariant to their length [**speech_and_language**]. The metric is identical to the cosine of the angle between the vectors v and w , as seen in Equation 2.4. The cosine similarity is by far the most commonly used similarity metric.

$$\text{cosine}(v, w) = \frac{v \cdot w}{|v||w|} = \frac{\sum_{i=1}^N v_i w_i}{\sqrt{\sum_{i=1}^N v_i^2} \sqrt{\sum_{i=1}^N w_i^2}} \quad (2.3)$$

$$\begin{aligned}
a \cdot b &= |a||b| \cos \theta \\
\frac{a \cdot b}{|a||b|} &= \cos \theta
\end{aligned} \tag{2.4}$$

2.3. Vision-Language Models

LLMs are inherently confined to processing exclusively text-based data. This limitation restricts their applicability in complex, real-world scenarios, where understanding and combining data from multiple modalities is crucial [vlm_frontier, qwen25vl]. Vision-language models (VLMs) are a class of models which respond to these limitations by combining visual and textual processing capabilities into a single architecture [vlm_frontier]. These models find applications involving both the comprehension and generation of multi-modal content, such as image captioning, and visual question answering [vlm_frontier].

2.3.1. Bounding Boxes

Bounding boxes represent the most fundamental method for annotating the position of an object within an image. A bounding box is the smallest rectangle that fully encloses the shape of the object [shape_analysis]. These boxes are defined within the image’s coordinate system, with its origin typically positioned at the top-left corner of the image [dive_into_dl]. The x-axis extends horizontally from this point, while the y-axis extends vertically. Coordinates can either be expressed in absolute pixel units or as normalized fractional values relative to the dimensions of the image. For this study, we focus exclusively on horizontal bounding boxes, which are aligned to the horizontal axis, also known as Feret Boxes [shape_analysis]. There are multiple formats for representing bounding boxes, with the left-top-right-bottom (LTRB) notation, which denotes the coordinates of the top-left and bottom-right corners of the bounding box, being a prominent option [dive_into_dl].

2.3.2. Intersection over Union

According to the definition from **object_detection**, the intersection over union (IoU) between two bounding boxes BB_a and BB_b is defined as described in Equation 2.5. The IoU can take on any value between 0 and 1, where a value of 0 means that there is no overlap between the two bounding boxes, and a value of 1 means that the two bounding boxes are identical. In the context of object detection, IoU is commonly used to evaluate the accuracy of predicted bounding boxes against ground truth bounding boxes [object_detection].

$$IoU(BB_a, BB_b) = \frac{\text{Area of intersection of } BB_a \text{ and } BB_b}{\text{Area of union of } BB_a \text{ and } BB_b} \tag{2.5}$$

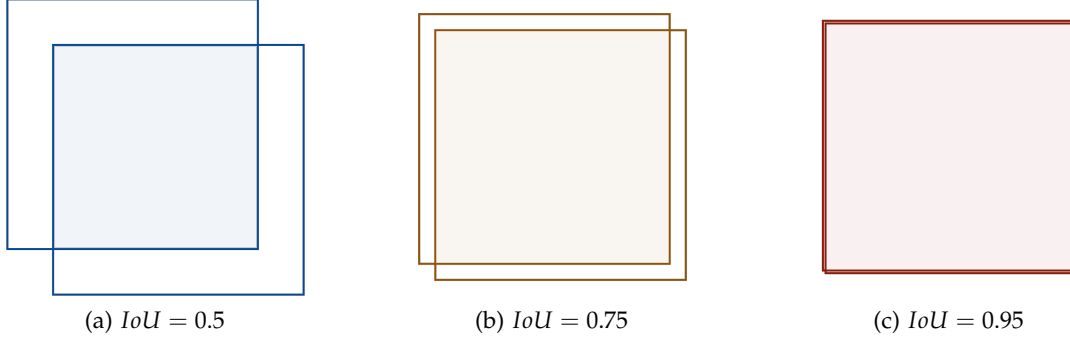


Figure 2.1.: A pair of identical bounding boxes at different IoU values. The shaded area is the intersection of the bounding boxes. As the IoU value increases, the area of the intersection approaches the area of the bounding box.

2.4. Retrieval-Augmented Generation

While LLMs have extensive general domain knowledge due to their enormous corpora of training data, compiled from various open-domain sources [**impact_dataset_rag_multi_hop**], they struggle with tasks that require domain-specific knowledge which they did not encounter during training [**rag_survey**]. This can lead to “hallucinations” and inaccuracies, as the model tries to synthesize a matching answer based on its domain-wise irrelevant training data [**rag_survey**, **chunk_size_effect_on_rag**]. RAG addresses this limitation, extending the usage of LLMs to applications requiring extensive knowledge in a specific domain [**rag**]. This is achieved by retrieving information from an external knowledge source comprised of application-relevant text passages, supplying additional context to the LLM during answer generation [**rag**, **rag_survey**].

2.4.1. Architecture of RAG Systems

While there are many advanced and extended versions of RAG systems, for this study we will focus on the standard Naive RAG architecture as depicted in Figure 2.2 [**rag_survey**]. Naive RAG is based on the original RAG architecture proposed by **rag**. Naive RAG systems consist of two modules:

Retriever: The retriever module consists of a query encoder and an external knowledge base [**rag**]. It is responsible for retrieving relevant context from the knowledge base, based on the user’s query [**rag_survey**]. The module is based on the bi-encoder architecture, with the query encoder q and document encoder d encoding texts into a shared embedding space [**rag**, **dual_encoders**]. The knowledge base is a vector database consisting of application-specific text passages z . Each passage is stored in the database as a vector embedding $d(z)$, encoded through the document encoder d [**rag**]. To identify the relevant passages for a query x , x is first transformed into a vector embedding $q(x)$ using the retriever’s query

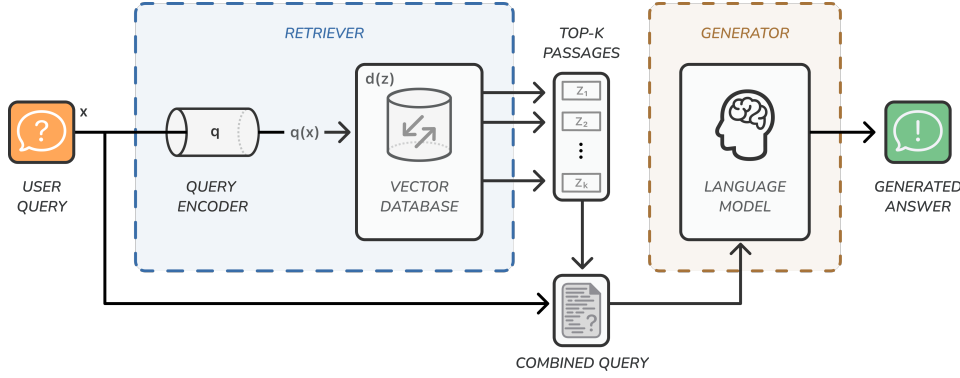


Figure 2.2.: Architecture of the Naive RAG system.

encoder [rag_survey]. Based on the similarity scores between the query embedding and the stored chunk embeddings, the top- k documents z with the highest similarity scores, are then retrieved from the database [rag, rag_survey].

Generator: The generator module is responsible for synthesizing the final answer based on the user’s query and the passages retrieved by the retriever [rag]. Firstly, the original query x and the retrieved passages z are combined into a single input query [rag_survey]. The LLM is then tasked with generating the final answer y , conditioned on this combined input [rag_survey, rag].

2.4.2. Indexing

In order to apply the RAG paradigm to knowledge-intensive tasks in a specific domain, the external knowledge base needs to be created from relevant data sources. This process is called Indexing [rag_survey]. Indexing begins with the preparation of the data sources into short text passages [rag_survey]. For the purpose of this study we will refer to this process as document segmentation. Document segmentation includes both DP, the conversion of unstructured documents, such as PDFs and images, into structured data [parsingunveiled], as well as chunking, the splitting of this data into smaller text passages called chunks [rag_survey]. Chunking is a necessary step for RAG systems, as both LLMs and encoders are limited in the number of tokens that fit into their context window [rag_survey]. Furthermore, indexing includes the encoding of these chunks into vector embeddings. Both the embeddings and the original chunks are then stored as key-value pairs in a vector database, allowing fast and frequent searches during retrieval [rag_survey].

The quality of the index construction has a crucial effect on the resulting RAG system [rag_survey]. It determines both the likelihood of retrieving relevant context as well as the quality of the generated answer. Especially chunking, which is often overlooked and seen as solely a technical requirement, has been found to be crucial for enhancing the quality of the knowledge base [rag_survey].

2.4.3. Source Attribution

Source attribution is a mechanism that provides transparency and traceability to the output of the RAG system by linking the generated text to their source documents [visa, llm_citations]. This allows the user to verify the LLMs claims by examining the provided sources [llm_citations]. Source attribution can be performed at different granularity levels. Document level source attribution provides citations to the entire documents that the retrieved passages belong to [visa, llm_citations]. While this approach enables the necessary verifiability, it introduces additional strain to the user, who has to find the relevant passages in the document [visa]. This effect is amplified for longer documents, such as CPGs. In order to mitigate this issue, recent research has suggested the concept of visual source attribution [visa]. Visual source attribution revolves around visual confirmation for the exact location of the retrieved information [visa]. This is achieved by highlighting the exact region of the retrieved text inside of the document [visa]. The position of the retrieved passage is therefore immediately visible to the user, making source attribution easy and seamless.

2.5. Document Parsing

Also known as document content extraction, DP aims to convert unstructured and semi-structured documents into structured, machine readable data formats [parsingunveiled, omnidocbench]. During this process elements, such as headings, tables, and figures, are extracted from the document while preserving their structural relationships. DP is crucial for many document-related tasks, providing access to previously unavailable information sources. Especially for LLMs, where leveraging additional training data is crucial for enhancing the model’s factual accuracy and knowledge grounding, DP plays an important role [omnidocbench, mineru]. With the emergence of the RAG paradigm, DP has also been critical in the creation of the knowledge database, as important information is often stored inside file formats which can not directly be processed by machines [docling]. While DP is used for converting a range of document formats into machine-readable content, we will focus solely on the parsing of PDF documents for the purposes of this thesis, as this is the data type that the oncology guidelines are stored as.

Converting PDF documents is particularly challenging due to their variable formatting, lack of standardization and focus on visual characteristics [docling]. The format not only includes born-digital files but also includes photographed and scanned documents. Therefore, DP systems need to be able to adapt to a wide range of different layouts, image qualities and document types, such as academic papers, invoices, or presentation slides [omnidocbench, intelligent_doc_parsing]. While there are many tools and implementations available for DP [docling, mineru, mineru_vlm, unstructuredio], most of them can be categorized into either modular pipeline systems or end-to-end VLM models.

2.5.1. Modular Pipeline Systems

Modular pipeline systems employ various different modules in a sequential order to perform DP. This modular design enables the targeted optimization of individual components and flexible integration of new modules and techniques [monkeyocr]. Additionally, by making use of lightweight models and integrating parallelization, pipeline systems can reach efficient parsing speeds [omnidocbench]. While different formations are possible, most implementations consist of three different stages [parsingunveiled].

Document Layout Analysis (DLA): According to parsingunveiled, DLA refers to the identification of the structural elements of a document, such as paragraphs, section headers, tables, figures, and mathematical equations, as well as their respective bounding boxes [parsingunveiled, docling_heron]. There are two types of methods for performing DLA. Uni-modal methods focus purely on visual features of the document in order to identify structural elements [parsingunveiled, pp_doclayout]. Notably, convolutional neural networks (CNN)- and transformer-based methods adapt models initially designed for object detection tasks, such as the YOLO [yolo] and DETR [detr] families of models, to accurately identify structural elements in document images [parsingunveiled, docling_heron]. Hereby, transformer-based methods excel at capturing global relationships between structural elements at the cost of computational intensity and expensive pre-training [parsingunveiled]. The second type of DLA methods are multi-modal methods. In addition to the visual representations, multi-modal methods also make use of the content and position of the pages' textual elements, performing DLA using a VLM [pp_doclayout, layoutlm_v3]. This approach allows more granular classifications and the analysis of highly complex layouts [parsingunveiled, pp_doclayout].

Content Extraction: To extract the content of the identified structural elements different recognizers are applied to the element regions based on their classifications [parsingunveiled, mineru, docling]. For textual elements, such as paragraphs or section headings, the textual content is identified using optical character recognition (OCR). OCR engines use techniques from computer vision in order to identify and extract text from images [parsingunveiled, ocr_survey]. Popular OCR engines include EasyOCR [easyocr] and the Tesseract OCR engine [tesseract]. In addition to extracting content using OCR, DP implementations often provide specific recognizers for additional element types [parsingunveiled, mineru]. Most commonly this includes a specific model for table structure recognition, referring to the extraction of table content into structured file formats, such as Hypertext Markup Language (HTML), extensible markup language (XML) or Markdown [parsingunveiled, docling, mineru, unstructured_open_source]. Other options for class-specific recognizers include mathematical formula recognition and chart recognition [mineru, mineru_vlm, parsingunveiled].

Relation Integration: During relation integration the identified elements are combined into the final output format. During this stage, rule-based methods and specialized AI models may

be employed, for example to filter out duplicate or unwanted elements or correct the reading order of the document [**parsingunveiled**, **mineru**, **docling**]. Depending on the chosen output format, this process might lead to the loss of information, such as the loss of bounding box information for an output in Markdown format [**docling_toolkit**].

Systems that follow the modular pipeline approach also have some inherent drawbacks. Mainly, due to handling the parsing of each structural element independently of each other, pipeline systems fail to capture information about the global context of the document, leading to semantic loss [**intelligent_doc_parsing**]. Additionally, because of the sequential nature of the pipeline approach, errors from different stages propagate through the pipeline [**intelligent_doc_parsing**, **mineru_vlm**].

2.5.2. End-to-End VLM models

Due to recent advancements in VLM architectures, end-to-end VLM models have emerged as a promising alternative to traditional pipeline-based approaches. Research, such as the General OCR Theory (GOT), have demonstrated the ability of VLMs to perform high accuracy OCR while being able to extract the content of tables, charts, or mathematical formulas using a singular model [**general_ocr_theory**]. Contrary to pipeline-based methods, VLM-based approaches are able to generate structured outputs directly from the input document, addressing the error propagation problem of modular pipelines [**monkeyocr**]. Additionally, these models demonstrate advantages in understanding the structure and hierarchy of complex documents [**parsingunveiled**]. VLM-based approaches can be divided into two further subcategories:

General-Purpose VLMs: General purpose VLMs are not trained exclusively for document-centric tasks, but are still able to show promising results for DP, due to their large parameter count and extensive training data [**mineru_vlm**, **qwen25vl**]. However, these models are often either proprietary or require extensive computational resources [**mineru_vlm**]. Additionally, they often struggle with documents that follow more complex layouts or contain densely packed text blocks [**mineru_vlm**].

Domain-Specific VLMs: Domain-specific VLMs are trained and optimized specifically for DP [**mineru_vlm**, **parsingunveiled**, **intelligent_doc_parsing**]. In recent years, there has been promising developments towards domain-specific VLMs that encapsulate DLA, content extraction and relation integration into a single model [**dotsocr**]. These models are able to achieve state-of-the-art performance on DP benchmarks, while being a fraction of the size of general-purpose VLMs [**dotsocr**, **mineru_vlm**]. As VLMs are not bound to the stages of traditional pipeline systems, there has also been additional research regarding models that are optimized for the direct generation of content-only outputs, most notably Markdown [**intelligent_doc_parsing**]. However, this approach inherently leads to the loss of information, such as positional information for the extracted elements, which is not included in the

Markdown format, making this class of models unsuitable for the purposes of this research [**dotsoocr**, **intelligent_doc_parsing**].

Recently, there has also been research towards multi-stage VLM-based approaches [**mineru_vlm**, **monkeyocr**]. These models use one or more VLMs in multiple stages, aiming to combine the computational efficiency of pipeline approaches with the improved accuracy and structure understanding of VLM-based methods [**mineru_vlm**]. However, especially when multiple VLMs are in use, these approaches come with a further increase in complexity and computational requirements and may show decreased performance in tasks such as reading order inference compared to single-stage VLM-based approaches [**mineru_vlm**, **dotsoocr**]. Current challenges regarding the development of VLM-based approaches are the risks of “hallucinations”, especially on longer documents [**mineru_vlm**, **docling_toolkit**], as well as their high computational requirements compared to modular pipeline systems [**docling_toolkit**].

2.6. Chunking

Chunking refers to the splitting of documents into small atomic units of information called chunks [**chroma_eval**, **rag_survey**]. While the term is directly linked to the recent emergence of the RAG paradigm, the underlying task of text division is fundamentally aligned to the established concept of passages in passage-based document retrieval [**passage_based_retrieval**, **passage_94**].

Despite the rapid adoption of RAG, the chunking process lacks a robust scientific taxonomy. Much of the terminology associated with modern chunking strategies originates from non-scientific sources, such as technical blogs, software documentation, and community tutorials. We find that the established taxonomy of passage-based document retrieval aligns with the types of modern chunking strategies. To ensure scientific stability, we therefore adopt the terminology proposed by **passage_94**. Specifically, **passage_94** categorizes passages into three distinct types: window passages, semantic passages, and discourse passages.

2.6.1. Window Passages

Window passages are determined by splitting the content of the document into parts of a fixed length. While in passage-based retrieval, length typically referred to the number of words in a passage [**passage_94**], with the advent of chunking the focus shifted towards measuring the number of tokens [**chroma_eval**]. Modern chunking strategies have further extended this method through sliding-window approaches, which introduce a fixed overlap between neighboring chunks to preserve contextual continuity [**llamaindex_chunking**]. These strategies provide a simple and computationally efficient way to perform chunking [**semantic_chunking**]. However, they disregard the content of the document, which may result in chunk borders appearing inside a single word or sentence [**chunking_comparison**].

2.6.2. Semantic Passages

Semantic passages aim to enhance retrieval quality by aligning passage borders to identified subtopics of the document [chunking_comparison]. However, they introduce significant additional computational complexity and may vary drastically in length [semantic_chunking]. Strategies from this category stem from the field of text segmentation, referring to “the task of dividing text into segments, such that each segment is topically coherent, and cutoff points indicate a change in topic” (p.1) [text_segmentation]. In recent years there have also been novel strategies proposed for this task that leverage LLMs to determine semantically independent chunks [lumberchunker].

2.6.3. Discourse Passages

Discourse passages are defined by the inherent structure of the document, such as sections, sentences, and paragraphs. Typically, these strategies recursively divide the document with increasing granularity until resulting chunks satisfy a specified maximum length constraint [langchain_splitting_recursively]. While the documents in passage-based document retrieval are simple unstructured text streams [passage_94], modern chunking techniques often process data in structured formats such as JavaScript Object Notation (JSON) or XML [docling_toolkit], especially when combined with DP. Recently, specialized strategies have emerged that leverage additional metadata from these formats, such as hierarchical relationships between elements, to produce chunks that follow the structure of the document more closely [docling_toolkit].

2.6.4. Metadata Attachments

In addition to their textual content, chunks can be enriched with metadata information [rag_survey]. This metadata can include information about the original document, such as its author, title, or publishing date. This enables the filtering of retrievable data based on document attributes, such as limiting the retrieval to documents published in a specific time frame [rag_survey]. Metadata attachments are also critical for providing source attribution. While document information enables traceability at the document level, additional metadata such as the page number and bounding box of the chunk achieves more granular grounding.

3. Methodology

3.1. Pipeline Overview

In order to compare different DP implementations and chunking strategies against each other, we developed a modular document segmentation pipeline. The pipeline's architecture, illustrated in Figure 3.1, follows a two-stage process. Firstly, raw PDF documents are transformed into a structured data format. This data is then partitioned into metadata-enriched chunks. The core principle of the pipeline's design lies in its modularity, allowing the seamless interchange of both the used DP implementation and the chunking strategy, while maintaining a unified interface for both modules.

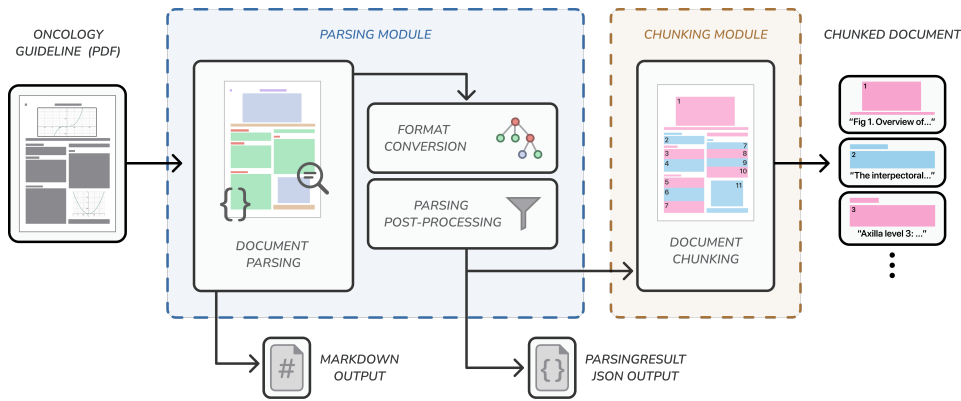


Figure 3.1.: Architecture of the document segmentation pipeline.

Parsing module: The parsing module is the first step of the pipeline. It integrates eight different DP implementations into a unified interface, normalizing their output formats into a standardized data format. Firstly, document elements and their structural information are extracted from the document using the underlying DP implementation. The normalized output then undergoes further post-processing steps, such as the filtering of unwanted element types. Finally, the result of the parsing operation is persisted to the file system as both a lossless JSON serialization and a lossy Markdown serialization for further processing and evaluating.

Chunking module: The chunking module is responsible for the splitting of the structured data into smaller chunks using one of multiple available chunking strategies. We adapt four established strategies to operate on the data format provided by the parsing module.

Additionally, we propose a novel approach to the chunking paradigm, enabling traceability of the chunk’s content on the token level. This allows for the determination of more accurate chunk bounding boxes, enabling high granularity, visual source attribution for downstream RAG applications.

3.2. Data Representations

A significant challenge for the evaluation and comparison of different DP implementations is the lack of standardization. As of the time of writing, every DP implementation defines their own data types, making direct comparisons very complex. In order to consolidate multiple different DP implementations into our modular pipeline and evaluate them against each other, we define our own universal data types to be used for the document segmentation process. Before further processing, the outputs of each DP implementation are first transformed into these data types.

3.2.1. ParsingResultType

A central issue arising from the fragmented methodologies of different DP methods is their lack of a universal terminology for recognized element types. For example, a paragraph gets classified as `NarrativeText` by the Unstructured.io framework [`unstructuredio`], while Docling [`docling`] names the same category as simply `TEXT`. Additionally, some methods provide classifications, which are not provided by others. One example for this is the addition of a `ref_text` element type in the MinerU implementation [`mineru`, `opendatalab`], referring to an entry in a bibliography. To address these issues, we aggregate all element categories from the evaluated DP implementations into a single collection, normalizing their categories into a unified terminology. We then provide mappings for each of the implementations to our universal categories. The full list of available `ParsingResultTypes` is provided in Table A.

3.2.2. ParsingBoundingBox

`ParsingBoundingBox` (Figure 3.2) serves as the fundamental data type to denote the location of an entity in the document. It expands upon a bounding box in LTRB format through the addition of a page number to support multi-page documents. The coordinates are stored as normalized fractional values of the page dimensions. Additionally, the data type includes the recursive attribute `spans`, which enables the assignment of bounding boxes of higher granularity, such as individual text lines.

3.2.3. ParsingResult

Inspired by the data structures of multiple DP implementations, such as Docling [`docling_toolkit`] and Google Document AI LayoutParser [`layout_parser`], we choose a tree structure to represent the output of the parsing module. This approach has the benefit of being able to model the structure and hierarchies of the structural elements through parent-child relationships,

```
class ParsingBoundingBox:
    page: int
    left: float
    top: float
    right: float
    bottom: float
    spans: list[ParsingBoundingBox]
```

Figure 3.2.: Python implementation of the ParsingBoundingBox data type.

ensuring a lossless representation of the original document. The ParsingResult data type (Figure 3.3) represents a node inside of this tree structure.

```
class ParsingResult:
    id: str
    type: ParsingResultType
    content: str
    geom: list[ParsingBoundingBox]
    parent: ParsingResult | None
    children: list[ParsingResult]
    metadata: dict
    image: str
```

Figure 3.3.: Python implementation of the ParsingResult data type.

The data type encapsulates all attributes identified during DP. Its classification and bounding box, which are identified through DLA, are stored in the type and geom fields respectively. The latter also permits multiple bounding boxes for a single structural element, to allow for more flexibility regarding the localizations returned by the DP implementation. The element’s content, identified during content extraction, is stored in the content field. Some implementations also persist images of figures or tables to the file system during content extraction, with image containing their respective paths. The id contains a document-wide unique identifier for the ParsingResult node. Lastly, parent and children model the tree structure.

The root node of the ParsingResult tree structure contains additional metadata about the parsing process, such as the elapsed parsing time, the used DP implementation, or the path to the parsed PDF document, in the metadata field. Traversing the tree from the root node in a depth-first manner iterates through the elements in reading order.

3.2.4. ChunkingResult

The ChunkingResult (3.4a) is the final output of the document segmentation pipeline. It provides a wrapper around the list of generated chunks, adding a metadata field for infor-

mation about the input document and the document segmentation process. Hereby, the `ChunkingResult` combines both information about the chunking process, such as the chosen chunking strategy, as well as the metadata from the root node of the preceding `ParsingResult` tree.

<pre>class ChunkingResult: chunks: list[Chunk] metadata: dict</pre>	<pre>class Chunk: id: str content: str metadata: dict geom: list[ParsingBoundingBox]</pre>
(a)	(b)

Figure 3.4.: Python implementation of the `ChunkingResult` (a) and `Chunk` (b) data types.

3.2.5. Chunk

The `Chunk` (3.4b) represents a singular passage used for the creation of the knowledge base in downstream RAG applications. In addition to the textual content of the chunk, which is stored in `content`, the data type contains the `ParsingBoundingBoxes` required for visual source attribution. Lastly, the `metadata` field contains additional information about the chunk, such as its token length.

3.3. Parsing Module

The parsing module revolves around extracting the content of a raw PDF file as a tree of `ParsingResult` nodes. It serves as an abstraction layer on top of various available DP implementations, unifying different DP approaches and output formats into a common interface. The module processes incoming documents through four sequential steps.

DP interaction: In the first step, the module handles the interaction with the underlying DP implementation. This includes request construction, preparing the input document in the required format of the implementation, and error handling. This logic is encapsulated through the abstract `_parse` function, extracting the result of the DP operation in the implementation-specific data structure.

Format conversion: Converting the implementation's data structure into the `ParsingResult` tree structure is the most crucial step to facilitate direct comparisons between different DP approaches. Due to the significant variations in implementations' data structures, this step is the most complex task of the parsing module, with its specific inner workings and required steps being highly dependent on the implementation's data representations. Typical steps include the transformation of the element's bounding box coordinates into the normalized

coordinates of the `ParsingBoundingBox`, the normalization of the elements classification into `ParsingResultType`, and the modeling of recognized hierarchical relationships in the `ParsingResult` tree. The abstract `_transform` function encapsulate this implementation-specific conversion logic.

Parsing post-processing: Even after the normalization to the universal `ParsingResult` tree structure, there are still some inherent differences between the outputs from the different DP approaches. In order to mitigate these differences and prepare the data for the chunking module, various rule-based post-processing steps are performed on the `ParsingResult` tree.

1. **Element filtering:** Most documents contain textual information that does not belong to the main content of the document, such as page numbers and repeating page headers or footers. Some DP implementations, such as MinerU [`mineru`], already remove these elements in their own post-processing stages. We remove any elements that belong to non-main content element types as well as any textual elements with empty content. Specifically we remove all elements from the following types: `[REFERENCE_LIST, REFERENCE_ITEM, PAGE_FOOTER, PAGE_HEADER, FORM_AREA, WATERMARK]`. This reduces structural bias in the comparison between DP implementations while removing unneeded information ahead of the chunking phase.
2. **Hierarchy inference:** In order to represent the document’s hierarchy as a tree structure, the relationships between the different elements need to be established. However, many of the evaluated DP implementations do not return their output in a tree structure directly and instead provide a list of document elements in reading order. Other implementations, such as Docling [`docling`], contain some hierarchy, such as the relationships between tables and their constituent table cells, while missing the relationships between section headings and the content belonging to their section. Inspired by the section heading matching process employed by Docling’s HybridChunker [`docling_toolkit`], we use the reading order of the document as well as the identified levels of the section headings to identify relationships between section headings and the nodes belonging to the section’s content. This process is crucial in order to fully model the inherent structure of the document, which is a central prerequisite for enabling the creation of discourse passages in the chunking module.
3. **Span-level bounding box identification:** In order to provide visual source attribution on a granularity higher than the `ParsingResult` level, more granular bounding boxes are needed. We use PyMuPDF [`pymupdf`], a Python library for the extraction and analysis of data from PDF documents, in order to extract the bounding boxes of individual lines of text inside of the bounding boxes of each `ParsingResult`. PyMuPDF enables the extraction of these bounding boxes either directly from the programmatic information contained in the PDF file or through the use of OCR. While we experimented with the use of both, due to the born-digital nature of the oncology guidelines, OCR did not show any improved accuracy while being substantially slower than programmatic text extraction. If lines contain excessive horizontal whitespace, PyMuPDF tends to split

them into separate bounding boxes. We address this by merging span bounding boxes if their vertical overlap is larger than a set threshold, resulting in unified bounding boxes for each line. For each element the identified span bounding boxes are stored in the span field of their respective ParsingBoundingBox.

Persistence: To enable the evaluation of the output quality of the DP implementations, the tree structure is serialized and persisted to the file system. Particularly, this includes two distinct serializations. Firstly, the content of the document is persisted as a lossy serialization to Markdown format. This format is used specifically for benchmarking the quality of the content extraction and is extracted directly from the implementation’s data structure through the `_get_md` function to ensure optimal adherence to the Markdown syntax. Some DP solutions require additional processing for this extraction. The second persisted file contains the lossless JSON serialization of the ParsingResult tree. Before serialization, metadata about the parsing process is added to the root node of the tree. This file is particularly important for evaluating the DLA capabilities of the DP implementations.

Function	Input	Output
<code>_parse</code>	PDF document	Custom data format
<code>_transform</code>	Custom data format	ParsingResult
<code>_get_md</code>	Custom data format	Markdown string

Table 3.1.: Overview over the abstract functions of the parsing module. Custom data format refers to the data types used by underlying DP implementation.

We provide integrations for eight different DP implementations, which we will introduce in the following sections. However, the core principle of the parsing module lies in its extendibility. In order to incorporate an additional implementation into the parsing module, the abstract functions described in Table 3.3 need to be implemented.

3.3.1. Unstructured.io

Unstructured.io is a prominent provider for DP, offering both a cloud-based API as well as an open-source library. For our study, we will focus on the open-source library version of Unstructured.io [`unstructuredio`, `unstructured_open_source`]. While the developers themselves explicitly highlight that the open-source library is not suited for large-scale production environments [`unstructured_open_source`], its inclusion within the documentation of popular RAG frameworks, such as Langchain [`langchain`] and LlamaIndex [`llamaindex`] make it a popular choice for a first point of contact with DP. Therefore, we will regard the open-source library as a baseline for the compared implementations. Unstructured.io follows a modular pipeline approach. Specifically, the implementation uses YOLOX, a uni-modal vision transformer, to perform DLA [`unstructured_open_source`, `yolox`]. The library also includes a specialized model for table structure recognition [`unstructured_open_source`].

3.3.2. Docling

Docling, which was developed by IBM in 2022, is one of the most popular available open-source DP libraries [**docling**, **docling_toolkit**]. Docling particularly stands out from other DP implementations through its permissive MIT license. To achieve this, Docling relies primarily on custom models instead of using third-party software, which is often not as permissive [**docling**]. Docling offers two different approaches for DP:

Parsing pipeline: Docling’s processing pipeline consists of three components: a PDF backend called DoclingParse, an internal model pipeline containing multiple AI models, and a post-processing stage [**docling_toolkit**, **docling**]. Firstly, the PDF backend extracts useful information from the document using both programmatic extraction as well as OCR techniques. This includes bounding boxes for every text element inside the document. The internal model pipeline then performs both the DLA as well as content extraction steps. Hereby, Docling provides their own models for table structure recognition with the TableFormer model [**docling_table_former**] as well as for DLA with their Heron model [**docling_heron**]. Heron is derived from RT-DETR [**rt_detr**], a uni-modal vision transformer, and retrained on DocLayNet [**doclaynet**], Docling’s own dataset for DLA. During DLA, identified bounding boxes are compared and intersected with bounding boxes retrieved from the PDF backend in order to provide more accurate localization [**docling_toolkit**]. Using TableFormer, Docling is the only evaluated open-source system, that provides individual content and bounding boxes for table cells. During post-processing the recognized elements are then combined into the DoclingDocument data type [**docling_toolkit**].

Granite Docling: Granite Docling is an end-to-end VLM for DP. It belongs to the group of domain-specific VLM models, specifically build for document understanding and conversion [**granite_docling**]. The model is very compact, consisting of around 258 million parameters [**smol_docling**, **granite_docling**]. With this model, Docling proposes the DocTags data format, a structured data format designed for representing both text and structure of the document through XML-style tags [**smol_docling**].

3.3.3. MinerU

Another popular choice for open-source on-device DP is the MinerU framework. Similar to Docling, MinerU also offers both a pipeline as well as a VLM-based approach for DP [**mineru**, **mineru_vlm**, **opendatalab**].

Parsing pipeline: MinerU extends the traditional processing pipeline through a pre- and post-processing stage. In the pre-processing stage unprocessable files are filtered out and metadata about the document is extracted using the PyMuPDF library [**mineru**, **pymupdf**]. This metadata includes the language of the document, the document’s page dimensions and the identification of scanned documents [**mineru**]. The pipeline then uses models from the DP model library PDF-Extract-Kit for DLA and content extraction [**mineru**, **opendatalab**,

doclayout_yolo, unimernet]. The model used for DLA is a fine-tuned version of LayoutLMv3, a multi-modal model [**mineru, layoutlm_v3**]. For content extraction, special models for formula and table structure recognition are employed by the pipeline. During the final post-processing stage, overlapping elements are cleaned up, unneeded elements are filtered out and the reading order of the document elements is inferred using a segmentation algorithm [**mineru**]. MinerU’s pipeline system is the only evaluated implementation that includes span-level bounding boxes in its output, removing the need for their identification during post-processing.

VLM: With MinerU2.5, the implementation’s offerings were expanded by a multi-stage VLM-based DP approach. This approach employs a 1.2 billion parameter VLM to perform DP in a two-stage approach [**mineru_vlm**]. Firstly, the model is used to perform DLA on the document, identifying elements and their reading order. In the second stage, the same model is applied again on individual image crops of the page element and is tasked to extract the content from the crop [**mineru_vlm**].

3.3.4. Gemini 2.5 Flash

Gemini 2.5 Flash is a closed-source proprietary model developed by Google with strong multi-modal capabilities across text, vision and audio [**gemini_2.5**]. While Google offers a more capable model in the form of Gemini 2.5 Pro, we follow the sentiment from **mineru_vlm**, that DP tasks “typically exhibit relatively low dependency on large-scale language models” (p.7) and, based on both models similar results on various image understanding benchmarks [**gemini_2.5**], instead opt to rely on the cheaper, faster Gemini 2.5 Flash model for our study. Gemini 2.5 Flash belongs to the group of general-purpose VLMs and, due to its closed-source nature, is only accessible through an application programming interface (API). The Gemini family of models received additional training in order to provide improved accuracy on object detection and image segmentation tasks [**gemini_2.5, gemini_image_understanding**]. We follow the documentation provided by Google on harnessing Gemini’s image understanding capabilities [**gemini_image_understanding**] to formulate a prompt, that takes advantage of this additional training for the DP task. The full prompt is available in Listing A.1.

3.3.5. LlamaParse

LlamaParse is a cloud-based paid DP service from the makers of LlamaIndex, a popular framework for building RAG systems and workflows [**llamaindex, llamaparse**]. While there is no official information on the architecture used for the DP system behind LlamaParse, its marketing as a “GenAI-native document parser” [**llamaparse**] as well as the option to provide custom prompts to the service suggests that at least some of its functionality stems from a VLM.

3.3.6. Google Document AI LayoutParser

LayoutParser from Google Document AI is another cloud-based paid provider of DP services [`layout_parser`]. Contrary to other services such as Google Document AI's Enterprise Document OCR [`google_ocr`], LayoutParser has a strong focus on identifying the relationships between different page elements. As such, LayoutParser can recognize the level of section headings, infer the hierarchy between different elements and extract the content from individual table cells. LayoutParser follows a multi-stage pipeline approach to perform DP, but, as LayoutParser is a proprietary system, its exact architecture is unknown.

3.4. Chunking Module

The chunking module transforms the hierarchical `ParsingResult` tree into a sequence of chunks. We propose a novel solution aimed at increasing the traceability of the chunk content to its constituent `ParsingResults`, therefore enabling visual source attribution in the downstream RAG system.

Prominent implementations of chunking strategies, such as the ones found in the RAG frameworks LlamaIndex [`llamaindex`] and Langchain [`langchain`], treat the chunking process as a division of the textual content of the document, typically in the form of a Markdown representation. This approach severs the link between the text and its underlying structural elements, complicating source attribution. Novel solutions, such as Docling's hybrid and hierarchical chunkers [`docling`, `docling_toolkit`], improve upon this by associating the resulting chunk with a list of constructing structural elements. However, this inclusion is binary, with no distinction between partially and fully included elements. This results in bounding boxes that always contain the entire structural element, regardless of how much of its text is included in the content of the chunk. Furthermore, while these methods identify discourse passages based on the relationships between the `ParsingResult` nodes, the resulting chunks lack positional information from included section headings.

To address these limitations, we propose a token-centric architecture, enabling the traceability of the chunks content to its constituent `ParsingResults` on the token level. We argue that since LLMs and encoders operate on tokens rather than characters, the token serves as the atomic unit of textual content.

```
class RichToken:
    element_id: str
    token_idx: int
    token: int
    text: str
```

Figure 3.5.: Python implementation of the `RichToken` data type.

The key concept of our proposed approach lies in the introduction of the `RichToken` (Figure 3.5). This data structure contains both the textual content of the token as well as its

origin in the `ParsingResult` tree. The `RichToken` is linked to its `ParsingResult` node through its `element_id`, the document-wide identifier of the `ParsingResult`. Its position inside the element’s content is encapsulated in the `token_idx` field.

Chunk Token Identification: The chunking process begins with the traversal of the document tree and the identification of `RichToken` groups that make up the resulting chunks. The specific logic for grouping these tokens, and therefore the type of the returned passages, is determined by the specific chunking strategy. As the module iterates through the `ParsingResult` nodes, their content is transformed into a stream of `RichTokens` using the `all-MiniLM-L6-v2` sentence transformer [sbert] as a tokenizer. To preserve the structural boundaries of the document tree, newline delimiters are placed between `ParsingResult` nodes, acting as textual representations of the document’s structure. As the strategy traverses the tree, identified `RichToken` groups are sequentially emitted through Python’s `yield` functionality. In addition, to avoid creating chunks that are too big for the encoder’s context window, a limit N for the maximum amount of tokens per chunk can be set on the chunking module.

Chunk Assembly: As the generator yields the identified `RichToken` groups, they are consumed by the `Chunk` assembly phase, constructing the final `Chunk` objects. Through the grounding provided by the `RichTokens`, the system identifies the included token ranges for the constituent `ParsingResults`. If only a partial number of the node’s tokens are included in the chunk, only the relevant span bounding boxes are included in the chunk’s positional information. We identify these spans by approximating the line that a token lies in, assuming constant token density through the content of the `ParsingResult` node. We find that by including the preceding and following lines of this approximation, inconsistencies from this approximation can be effectively mitigated, while still providing bounding boxes at a high granularity. Finally, the content of the chunk is aggregated and metadata, such as the chunk’s token length, is stored in its respective field.

Our proposed chunking architecture enables precise visual source attribution for established chunking strategies, while providing structural information that the strategies can leverage during segmentation. We also address the limitations of previous visual source attribution systems [visa], by enabling attributions to span over multiple pages. We provide implementations for four commonly used chunking strategies, with the architecture of the chunking module allowing the integration of additional strategies for future experiments.

3.4.1. Fixed-Size Chunking

Fixed-size chunking implements the window passage approach. It splits the document into chunks of the chunking module’s maximum chunk length N , disregarding logical boundaries in favor of uniform chunk sizing [llamaindex_chunking]. The strategy traverses the `ParsingResult` tree in reading order, creating a queue of the document’s `RichTokens` in the progress. When the queue reaches a length greater than N , the queue’s first N tokens

are emitted and assembled into a chunk. In order to maintain some contextual continuity between the chunks, we implement a sliding-window mechanism with an overlap of O tokens. This mechanism can lead to improved recall during the retrieval phase of downstream RAG systems [**semantic_chunking**, **chunking_comparison**]. After the chunk is emitted, only the first $N - O$ tokens are removed from the queue, leaving O tokens to form the beginning of the subsequent chunk. After traversing the `ParsingResult` tree, any residual tokens are grouped together to form a final undersized chunk.

3.4.2. Recursive Character Chunking

Recursive character chunking leverages the structure of the textual content of the `ParsingResult` nodes to identify discourse paragraphs inside the document. The approach functions similarly to fixed-size chunking, however recursive character chunking utilizes an hierarchical list of delimiters (e.g., paragraphs, sentences, words) to define chunk boundaries [**chunking_comparison**, **langchain_splitting_recursively**, **langchain**].

The delimiters used in this implementation are adapted from `LangChain`'s `RecursiveCharacterTextSplitter` [**langchain**, **langchain_splitting_recursively**], with punctuation added to better identify sentence endings, as suggested by **chroma_eval**. This results in the following delimiters: `["\n\n", "\n", ".", "!", "?", "\u2014", ""]`.

Once the queue's length exceeds N , the strategy splits the tokens using the highest-order delimiter. If there still exists a split which is larger than N , the process recurses on the oversized split with the next delimiter in the list. This "coarse-to-fine" approach preserves logical groupings while avoiding unnecessary fragmentation [**chunking_comparison**]. Similar to fixed-size chunking, recursive character chunking also incorporates a sliding window approach with an overlap of O tokens between adjacent chunks.

3.4.3. Breakpoint-based Semantic Chunking

Breakpoint-based semantic chunking separates the document at the sentence level, inserting breakpoints in between sentences to denote chunk borders [**semantic_chunking**]. Instead of relying on structural markers, the strategy generates semantic passages by identifying shifts in the topics of the document.

As the strategy traverses the document tree, `ParsingResult` nodes are split into individual sentences using the `punkt` tokenizer from the Natural Language Toolkit (NLTK) [**punkt_tab**, **nlk**]. The strategy then computes their vector embeddings and the cosine distance of each adjacent sentence pair using the `all-MiniLM-L6-v2` encoder. A high distance, which is the inverse of the sentences similarity, denotes a topical shift between these sentences [**langchain_splitting_recursively**].

We then insert a breakpoint between sentences which have a higher distance than the Q -th percentile of all calculated distances. Breakpoint-based semantic chunking has no regard for the size of its produced chunks, leading to a large variability in chunk sizes. To mitigate this issue, we introduce a minimum chunk size M . If the strategy produces a chunk which is shorter than M , we combine it with the next splits until their combined length is greater than

M. To handle the inverse problem, we use the strategy of recursive character chunking in order to further split oversized chunks.

3.4.4. Hierarchical Chunking

Hierarchical chunking, which is inspired by the hybrid chunker from the Docling toolkit [**docling_toolkit**], generates discourse passages by leveraging the tree structure of the ParsingResult. The structure and hierarchy of the document are preserved in the final chunks by prepending relevant section headers to the chunk’s content.

In order to prevent deep hierarchical structures from taking up a large part of the final chunk, a token budget B_{headings} is imposed on the length S of the section heading tokens. If adding an additional heading causes S to exceed B_{headings} , the highest-level ancestors are removed from the list until the size constraint is satisfied.

As the algorithm traverses the document tree, it processes each ParsingResult node following a three-step logic:

1. **Subtree evaluation:** The algorithm determines the token count of the entire subtree rooted at the current node. If the length of the subtree is less than the remaining capacity, $N - S$, the subtree is grouped together into a single chunk. This prevents unnecessary fragmentation of small structures inside the document.
2. **Recursion:** If the subtree exceeds the size limit the node’s content is appended to the heading tokens and the algorithm recurses onto the children of the node. After recursion, adjacent children nodes are merged as long as they were not split any further during recursion and their combined length does not exceed $N - S$. This ensures that resulting chunks are as close to the length N as possible. After merging, the content’s tokens are prepended to each of the splits and the splits are returned.
3. **Leaf-splitting:** If the algorithm reaches a leaf node that exceeds the available space $N - S$, it results to using recursive character splitting to determine the chunk boundaries. Following the logic from the recursion step, the content’s tokens are prepended to the splits before they are returned.

3.5. Evaluation Framework

3.5.1. Document Layout Analysis Evaluation

The goal of the DLA evaluation is to assess the correctness of the bounding boxes and type labels produced by the parsing module [**icdar2009**]. While there are multiple datasets available for this task [**docbank**, **doclaynet**, **omnidocbench**], we will use the PubLayNet dataset [**publaynet**] for our evaluation. While many datasets focus on evaluating DLA on a range of different document types such as forms, invoices or handwritten documents, PubLayNet consists solely of medical scientific articles [**omnidocbench**, **publaynet**]. This format closely resembles the format of the oncology guideline documents which makes it a

suitable choice for this evaluation. Compromised of over 360.000 automatically annotated document pages collected from PubMed Central Open Access (PMCOA), PubLayNet is one of the largest datasets for DLA [**publaynet**]. As the dataset in its entirety is no longer publicly available and far too large for the purposes of this thesis, we will use **publaynet-mini**, a small subset of 500 pages of the original dataset for this evaluation [**publaynet-mini**]. As seen in Table 3.2, the subset contains around 5000 ground truth annotations for elements from 5 different classes.

Category	Annotations
Text	3,676
Title	1,000
List	73
Table	128
Figure	172
Total	5,049

Table 3.2.: Distribution of ground truth annotations across the different element types contained in the **publaynet-mini** subset of the PubLayNet dataset.

To assess the performance of different predictors on object detection tasks such as DLA, average precision (AP) is the most commonly used metric [**object_detection_survey**]. Previous evaluations of DLA models on the PubLayNet dataset also use a version of this metric [**icdar2021_competition**]. However, AP relies on the predictor’s confidence values, indicating how confident the predictor is about a predicted bounding box and class label. As most of the DP implementations provide “hard-predictions”, which do not contain any confidence values, the AP is not a viable metric for the purposes of this study [**lrp_error**, **eclair**].

For this reason, we will compare the implementations based on their achieved F1 score. Similarly to AP, this metric takes into account two important measures for object detectors: precision and recall [**object_detection_metrics**]. According to **object_detection_metrics**, “Precision is the ability of a model to identify only relevant objects. [...] Recall is the ability of a model to find all relevant cases [...]” (p. 9). In order to calculate their values, firstly the detected bounding boxes (DTBBs) are classified into true positives (TPs) and false positives (FPs). A DTBB is classified as a TP if there exists a ground truth bounding box (GTBB) from the same class, so that their IoU is greater than a given threshold. One GTBB can not be matched to multiple DTBBs. If there does not exist a GTBB that fulfils these criterions, the DTBB is classified as a FP. Any GTBBs which were not matched to a DTBB are classified as false negatives (FNs). Following the definition from **object_detection_metrics** for a model that, on a dataset with G GTBBs, outputs N DTBBs, out of which S , ($S \leq N$) are TPs, precision and recall can be formulated as shown in Equation 3.1 and Equation 3.2.

$$\text{Pr} = \frac{\sum_{n=1}^S \text{TP}_n}{\sum_{n=1}^S \text{TP}_n + \sum_{n=1}^{N-S} \text{FP}_n} = \frac{\sum_{n=1}^S \text{TP}_n}{\text{all detections}} \quad (3.1)$$

$$\text{Re} = \frac{\sum_{n=1}^S \text{TP}_n}{\sum_{n=1}^S \text{TP}_n + \sum_{n=1}^{G-S} \text{FN}_n} = \frac{\sum_{n=1}^S \text{TP}_n}{\text{all ground truths}} \quad (3.2)$$

The F1 score is the weighted harmonic mean between precision and recall and is calculated as defined in Equation 3.3 [**object_detection_metrics**]. The F1 score is calculated for a single class at a set IoU threshold. Selecting a higher threshold will lead to a stricter metric as predictions need to be more precise to be counted as a TP [**object_detection_metrics**]. A F1 score calculated at an IoU threshold T% is commonly referred to as F1@T [**scenescript**].

$$F_1 = 2 \frac{\text{Pr} \cdot \text{Rc}}{\text{Pr} + \text{Rc}} \quad (3.3)$$

For scenarios with multiple classes, such as the PubLayNet dataset, the Macro F1 score can be used to assess the overall performance of the predictor [**maximize_f1**]. The Macro F1 score is the mean of the single class F1 scores. For a dataset with M different classes, the calculation of the Macro F1 score is described in Equation 3.4. Hereby, $N_{:j}$ and $G_{:j}$ denote the DTBBs and GTBBs belonging to elements of class j [**maximize_f1**].

$$F_{1\text{Macro}}(N, G) = \frac{1}{M} \sum_{j=1}^M F_1(N_{:j}, G_{:j}) \quad (3.4)$$

The DP implementations will be evaluated on both their single-class and Macro F1 scores. Specifically, their (Macro) F1@50 and F1@50:95 will be compared against each other. F1@50:95 refers to the mean of the F1 values calculated at 10 evenly spaced IoU thresholds between 0.5 and 1.0 and is inspired by the primary challenge metric found in the MS COCO dataset [**coco**]. This rewards implementations, which provide more accurate bounding boxes [**coco**]. F1@50 is chosen, as a threshold of 50% is one of the most commonly used threshold values for metrics in object detection [**object_detection_metrics**]. To calculate these metrics, the `faster-coco-eval` package is used to determine the recall and precision values at the IoU thresholds [**faster-coco-eval**].

3.5.2. Content Extraction Evaluation

The goal of content extraction evaluation is to evaluate the quality of the extracted textual content from the document. In the context of oncology guidelines, where an error in the extraction of textual content can have a fatal effect, potentially altering clinical recommendations, assessing the quality of the content extraction is a critical aspect of our evaluation.

OmniDocBench has established itself as the leading benchmark for performing end-to-end evaluations [**omnidocbench**, **paddleocr**, **mineru_vlm**, **monkeyocr**, **dotsocr**, **docling_toolkit**], in particular for evaluating the quality of the DP implementation’s Markdown output. The benchmark includes nine different PDF document types, such as academic literature, newspapers, and financial reports, from 3 different language types, Chinese, English, and mixed [**omnidocbench**]. In total, the benchmark contains 1355 single-page PDF documents [**omnidocbench**]. For the purpose of this study, we limit our evaluation to english academic

literature. These documents closely resemble the layout of the CPGs, including both single and double-column layouts with complex tables and figures. After filtering, we are therefore left with a total of 129 PDF documents for our evaluation.

The dataset behind OmniDocBench was created through a semi-automatic process. Initial annotations are retrieved using LayoutLMv3 [`layoutlm_v3`] for DLA and PaddleOCR [`paddleocr`], UniMERNet [`unimernet`], and GPT-4o [`gpt4o`] for content extraction. Annotators manually refine the extracted annotations, correcting reading order and extracted content, as well as affiliating captions with their respective figures and tables. In a final steps expert researchers review and correct mathematical formulas and tables to ensure accuracy in the final annotations. In total, the dataset includes over 20,000 annotated structural elements.

The evaluation of a single DP implementation follows a three-step process:

Extraction: In a first step, the documents’ elements are extracted from the Markdown texts. Since the Markdown format is purely textual, this step is primarily performed using regular expression matching [`omnidocbench`]. The format of tables varies depending on the DP implementation with Markdown, HTML and \LaTeX formats being possible. To prevent interference between extraction steps, the extraction of different types follows a specific order. After extraction, Markdown tables are converted into HTML format for further processing. In total, five different element types are extracted: \LaTeX and HTML tables, display formulas, code blocks, and paragraphs [`omnidocbench`].

Matching: The extracted elements are now matched to ground truth elements of the same type through a process called Adjacency Search Match [`omnidocbench`]. First, the normalized edit distances between each possible pair is calculated, with pairs that exceed a specific threshold being considered as a successful match. As different implementations separate paragraphs at different positions, fuzzy matching is used to identify any paragraphs that are substrings of a respective matching partner. If so the substring paragraph is merged with its adjacent paragraphs until the pair’s normalized edit distance starts to increase [`omnidocbench`]. This ensures that the way paragraphs are separated does not influence the matching process. Some element types, such as headers and figure captions, are automatically removed by some implementations. To ensure a fair comparison, these elements are also removed before the calculation of the metrics [`omnidocbench`].

Metric calculation: OmniDocBench provides a multitude of different evaluation metrics for the extracted element [`omnidocbench`]. Based on the characteristics of the oncology guidelines, we select the following subset of metrics for our evaluation:

1. **Normalized Edit Distance:** We rely on the normalized edit distance to evaluate how well the DP implementation extracts content from the textual elements of the oncology guidelines. This is a crucial step for ensuring that clinical recommendations remain intact and unaltered.

The Generalized Levenshtein Distance (GLD), also known as the edit distance, is a metric that measures the textual similarity between two strings $X, Y \in \Sigma^*$, with Σ^* being the set of strings over an alphabet Σ [levenshtein, edit_distance]. $\lambda \notin \Sigma$ is the empty string. The metric denotes the minimum cost of transforming X into Y through weighted elementary edit operations. According to **edit_distance**, an “elementary edit operation $[T]$ is a pair $(a, b) \neq (\lambda, \lambda)$, often written as $a \rightarrow b$, where both a and b are strings of lengths 0 or 1.” (p.1) There are three elementary edit operations: insertions ($\lambda \rightarrow a$), substitutions ($a \rightarrow b$), and deletions ($b \rightarrow \lambda$). The edit transformation of X into Y , $T_{X,Y} = T_1 T_2 \dots T_l$, is a sequence of elementary edit operations. Using a weight function γ which assigns a nonnegative real number $\gamma(T)$ to each elementary edit operation T , the weight of the edit transformation $T_{X,Y}$ is computed as defined in Equation 3.5 [edit_distance].

$$\gamma(T_{X,Y}) = \sum_{i=1}^l \gamma(T_i) \quad (3.5)$$

Following the definitions from **edit_distance**, given $X, Y \in \Sigma^*$, the GLD is then defined as in Equation 3.6.

$$\text{GLD}(X, Y) = \min\{\gamma(T_{X,Y})\} \quad (3.6)$$

The GLD is not normalized with respect to the lengths of X and Y . A short string X that requires the same elementary edit operations as a longer string X' , $|X| < |X'|$ is therefore not additionally penalized. The normalized GLD, which is defined in Equation 3.7, addresses this issue. For a weight function γ , that assigns the same weight to insertions and deletions of any character, $d_{\text{N-GLD}}$ is a metric over Σ^* whose values are in $[0, 1]$ [edit_distance].

$$\alpha = \max\{\gamma(a \rightarrow \lambda), \gamma(\lambda \rightarrow b), a, b \in \Sigma\} \quad (3.7)$$

$$d_{\text{N-GLD}}(X, Y) = \frac{2 \cdot \text{GLD}(X, Y)}{\alpha \cdot (|X| + |Y|) + \text{GLD}(X, Y)}$$

2. **Tree-Edit-Distance-based Similarity (TEDS):** As complex tables are a common occurrence in the oncology guidelines, examining the quality of the table structure recognition of the different implementations, is a key requirement for our evaluation. TEDS measures the similarity between two HTML tables. In order to calculate this metric, extracted \LaTeX tables are first converted into HTML format [omnidocbench]. TEDS builds on top of the tree edit distance proposed by **tree_edit_distance**. Hereby, the tree edit distance is defined as “the minimum-cost sequence of node edit operations that transforms [a] tree F into [another tree] G ” (p.1) [tree_edit_distance].

In HTML, tables are represented as tree structures. The root node has two children, `thead` and `tbody`, grouping the table’s header rows and body rows respectively. Each

table row tr is made up of table cells td , the leaves of the table tree. A cell has three attributes. The attributes “colspan” and “rowspan” denote the respective number of columns and rows that the cell stretches across. Lastly, “content” includes the textual content of the table cell [teds].

TEDS defines three different operations and their respective costs. Inserting or deleting a node has a cost of 1. The cost of substituting a table node n_o with n_s varies on the type and attributes of the node. If one of the node is not a leaf node their substitution, such as switching the order of two rows, has a cost of 1. When both nodes are table cells, their substitution cost is 1 if their spanning columns or rows are different. If the leaf nodes differ in their content, their substitution cost is the normalized GLD between their contents [teds]. Following the definition from teds, the TEDS between two table trees F and G is then computed as in Equation 3.8. Hereby, $|T|$ denotes the number of nodes in a table tree T and EditDist is the tree edit distance. The value of TEDS is confined to $[0, 1]$ [teds].

$$\text{TEDS}(F, G) = 1 - \frac{\text{EditDist}(F, G)}{\max(|F|, |G|)} \quad (3.8)$$

In addition to the regular TEDS, OmniDocBench provides a structure-only version of this metric [omnidocbench]. This version disregards differences between the content of table cells, effectively setting the substitution cost of two table cells that span the same amounts of rows and columns to 0. In combination, these metrics help discern between the DP approach’s ability to understand the structure of the table and its ability to extract accurate textual information from the table cells.

3. **Normalized edit distance for the reading order:** Evaluating the reading order of the implementation’s output is an important validation step to ensure that the system is able to adapt to the double and single column layouts of the oncology guidelines. The normalized edit distance of the reading order is used for this evaluation [omnidocbench]. In order to create a fair comparison only text elements are included for this evaluation, as the placement of floating elements such as tables or figures can be somewhat ambiguous [omnidocbench]. Each of the ground truth elements contains an integer referring to their index in the natural reading order of the source document. The list of these indices is referred to as I . The indices are then ordered by the reading order of their matched predictions in the Markdown document. This list of reordered indices is referred to as I' and denotes the ordering as returned from the DP implementation. The normalized edit distance between I and I' is then calculated, with the elementary edit operations being the removal, insertion and substitution of a list element.

The chosen metrics and document filters are stored in a configuration file and passed to the benchmark during evaluation. The full OmniDocBench configuration file that is used for our evaluation is displayed in Figure A.1.

3.5.3. Chunking Evaluation

The quality of the DP process is meaningless if the relevant passages can not be found during the retrieval phase. As previously discussed, chunking can have a significant impact on the retriever’s ability to find the correct passages. However, measuring the quality of chunking is not trivial and generally overlooked during evaluation [**chroma_eval**]. Established RAG evaluation frameworks, such as Ragas [**ragas**], evaluate the quality of the retrieved context by using a LLM to decide whether a retrieved chunk is relevant to the query [**ragas**]. This approach not only introduces additional complexity and uncertainty but also fails to measure the ratio of relevant information inside the retrieved chunks.

chroma_eval propose a framework that focuses specifically on evaluating the influence of the chunking process on the retrieval phase. Their methodology builds on the fact that retrieving irrelevant tokens results in unnecessary computational strain and distractions during generation [**llm_context_distractions**, **chroma_eval**]. Therefore they propose a novel evaluation approach that evaluates the retrieved context on the token level. Their contributions include the proposal of specialized evaluation metrics as well as a framework for the creation of the evaluation dataset. We follow their evaluation approach while adapting it to our own data types.

Dataset creation: To compare the chunking strategies against each other, a corpus of oncology guideline documents needs to be selected for the evaluation. In addition, question-answer (QA) pairs are needed to evaluate the quality of the retrieval. Hereby the framework provides utilities for the generation of synthetic QA pairs from the document corpus through the use of a LLM. However, for our evaluation we use manually annotated QA pairs, created by medical professionals from the Technical University of Munich (TUM) university hospital. While this dataset is not available to the general public, we were permitted to use both the documents as well as the QA pairs for this evaluation. In total, the dataset contains **[(TODO: count about 30?)]** QA pairs created for a document corpus of **[(TODO: count)]** german oncology guidelines published by AWMF.

The documents are first processed by the parsing module to prepare them for the chunking strategies. Hereby, we choose the used DP implementation based on the results of the content extraction evaluation. The document corpus D contains the tokens across all of the parsed guideline documents. The parsed documents are then processed by the chunking module once for each chunking strategy. For each strategy i , the chunks across all of the documents are contained in the chunk corpus C_i . Each chunk $c \in C_i$ is a set of tokens such that $c \subseteq D$. For each question q , the answer subset $T_e, T_e \subseteq D$ contains the information that is relevant for answering q .

Evaluation: **chroma_eval** argue that a metric for measuring the quality of the retrieval phase should “take into account not only whether relevant excerpts are retrieved, but also how many irrelevant, redundant, or distracting tokens are [...] retrieved”. During the evaluation phase a retriever is created for each chunk corpus C and its performance is evaluated using the QA pairs.

Based on the question q , the retriever retrieves a set of chunks $c = \{c_1, c_2, \dots, c_l\}, c_i \subseteq D, c_i \in C$ from the chunk corpus C . $T_r, T_r = \bigcup_{c_i \in C} c_i$ is the set of all tokens contained in the retrieved chunks. As T_r is a set, it does not contain any duplicate tokens [discrete_math]. However, for many chunking strategies that employ a sliding-window approach, the same token might be included in multiple retrieved chunks. Retrieving the same token twice introduces additional noise, which should be penalized by the evaluation metrics.

For the metric calculation, **chroma_eval** account for this redundancy by noting that the cardinality of T_r includes the multiplicity of included tokens [chroma_eval]. To formulate this logic with the needed mathematical rigor, we define M_r as the multiset over T_r . $M_r : T_r \rightarrow \mathbb{N}_0$ is a function such that if $M_r(t) = k > 0, t \in T_r$, then t appears with multiplicity k in M_r [discrete_math]. The cardinality of the multiset is defined as $|M_r| = \sum_{t \in T_r} M_r(t)$ [discrete_math]. Through $|T_r|$ and $|M_r|$, we can express both the number of unique retrieved tokens and the number of total retrieved tokens, including duplications.

Following this principle, **chroma_eval** propose three distinct evaluation metrics. For each evaluated chunking strategy, the mean of each metric over the QA pairs is reported [chroma_eval].

1. **Token-wise precision:** Analog to its definition in the DLA evaluation, the token-wise precision measures the ratio of retrieved tokens that are TPs, meaning relevant for the question. Hereby, returning a relevant token twice leads to a lower precision. Given a query q and a chunked document corpora C , the token-wise precision is calculated as in Equation 3.9 [chroma_eval].

$$\text{Pr}_q(C) = \frac{|T_e \cap T_r|}{|M_r|} \quad (3.9)$$

2. **Token-wise recall:** Token-wise recall measures how many of the relevant tokens were successfully retrieved. Given a query q and a chunked document corpora C , the token-wise recall is then calculated as in Equation 3.10 [chroma_eval].

$$\text{Re}_q(C) = \frac{|T_e \cap T_r|}{|T_e|} \quad (3.10)$$

3. **Token-wise IoU:** Similar to the IoU between bounding boxes, the token-wise IoU calculates the overlap between the retrieved tokens and the ground truth tokens. Given a query q and a chunked document corpora C , The token-wise IoU is calculated as described in Equation 3.11 [chroma_eval].

$$\text{IoU}_q(C) = \frac{|T_e \cap T_r|}{|T_e| + |M_r| - |T_e \cap T_r|} \quad (3.11)$$

4. **Precision_Ω:** Precision_Ω refers to the precision value of a “perfect” retriever that always retrieves the set of chunks c_{opt} consisting of every chunk that contains tokens from T_e . The metric therefore provides an upper bound for the token efficiency given perfect recall [chroma_eval].

3.5.4. Evaluation Environment

All evaluations are performed on a 2023 MacBook Pro equipped with a M3 processor and 16 GB of unified memory. In order to optimize the performance of this hardware, especially for the locally deployed VLM-based DP approaches, we utilize Apple’s MLX engine [mlx] when possible.

4. Results

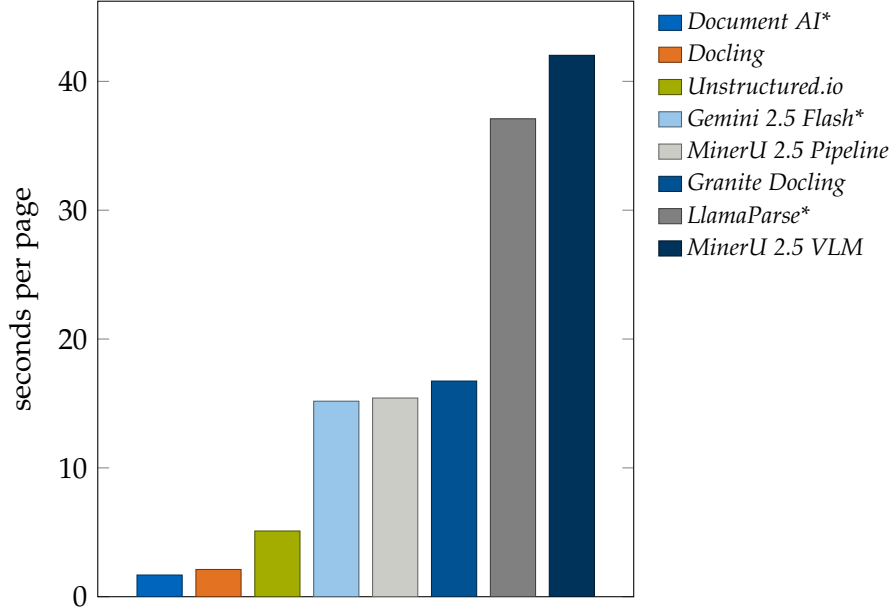
	text	title	list	table	figure	all
Unstructured.io	0.8123	0.8032	0.1269	0.9337	0.6138	0.6216
Docling	0.8687	0.8775	0.8022	0.9530	0.5951	0.8063
Granite Docling	0.6737	0.6309	<u>0.6329</u>	0.9001	0.1811	0.5296
MinerU 2.5 Pipeline	<u>0.8735</u>	<u>0.9558</u>	0.4771	0.9784	<u>0.6534</u>	0.6528
MinerU 2.5 VLM	0.9119	0.8822	0.5702	<u>0.9796</u>	0.2508	0.5941
LlamaParse	0.7711	0.6370	0.0000	0.6831	0.0000	0.4110
Document AI	0.7830	0.9789	0.0000	0.9911	0.9848	<u>0.7393</u>
Gemini 2.5 Flash	0.8242	0.7619	0.1271	0.8725	0.6530	0.6153

(a) F1@50

	text	title	list	table	figure	all
Unstructured.io	0.7583	0.6029	0.0682	0.8707	0.4987	0.5095
Docling	0.8206	<u>0.6311</u>	0.7406	0.9143	0.4952	0.6650
Granite Docling	0.6241	0.4302	<u>0.5694</u>	0.8497	0.1608	0.4382
MinerU 2.5 Pipeline	<u>0.8097</u>	0.6170	0.4331	0.9407	0.5436	0.5342
MinerU 2.5 VLM	0.8032	0.4461	0.4906	<u>0.9409</u>	0.2034	0.4338
LlamaParse	0.7240	0.3057	0.0000	0.6594	0.0000	0.3073
Document AI	0.7136	0.6595	0.0000	0.9669	0.9524	<u>0.6015</u>
Gemini 2.5 Flash	0.7347	0.5002	0.0760	0.7636	<u>0.5822</u>	0.4890

(b) F1@50:95

Figure 4.1.: F1 scores of the evaluated DP implementations on the PubLayNet dataset. (a) contains the F1@50 scores, (b) contains the F1@50:95 scores. Scores are reported per element type. The column all reports the respective Macro F1 score for each implementation. Highest values are bolded and smallest values are underlined. Higher values are preferred.



Method	Parsing		Transformation	
	mean	std	mean	std
Docling	<u>2.1146</u>	<u>1.3147</u>	0.0017	0.0024
Document AI	1.6844	0.4090	1.5988	0.2888
Gemini 2.5 Flash	15.1748	13.1830	1.8871	9.2539
Granite Docling	16.7379	63.5661	0.0017	0.0034
LlamaParse	37.0956	41.6822	0.7354	0.1782
MinerU 2.5 Pipeline	15.4215	20.2052	0.0020	0.0059
MinerU 2.5 VLM	42.0182	35.4030	<u>0.0017</u>	<u>0.0012</u>
Unstructured.io	5.1010	5.5998	0.0008	0.0005

Figure 4.2.: Mean and standard deviation of the DP implementation's parsing and transformation times per page. Times are reported in seconds. Fastest times are bolded and slowest times are underlined. Lower values are preferred.

	Text ^{Edit} ↓	Table ^{TEDS} ↑	Table ^{S-TEDS*} ↑	Read ^{Edit} ↓	Overall ↑
Unstructured.io	0.0836	64.3603	82.0875	0.1161	81.464
Docling	0.078	66.3294	85.2592	0.0791	83.5388
Granite Docling	0.1365	64.211	70.0415	0.089	80.5537
MinerU 2.5 Pipeline	<u>0.0439</u>	<u>80.2155</u>	<u>90.1323</u>	0.0386	<u>90.6538</u>
MinerU 2.5 VLM	0.0247	86.0939	92.8198	0.0059	94.345
Gemini 2.5 Flash	0.0455	65.3683	72.6077	0.0409	85.5765
Document AI	0.0452	62.935	74.8045	<u>0.0261</u>	85.2689

Table 4.1.: Results of the DP implementation on the OmniDocBench benchmark.

5. Discussion

6. Conclusion

- How to improve tables
- How to include figures
- Make bounding boxes more granular *to* token level instead of line level

A. General Addenda

Listing A.1: Prompt to apply the Gemini 2.5 Flash model to DP tasks. Gemini models are trained to output coordinates from 0 to 1000, with the origin at the left-top corner of the image. Additionally, they are trained to provide bounding boxes as tuples in the (y_0, x_0, y_1, x_1) format. In order to maximize the accuracy of detected bounding boxes, `box_2d`, the key used in Google's official documentation, is used to denote the bounding box tuples in the output JSON.

```
<system_role>
You are an expert Document Layout Analysis AI. Your goal is to perfectly transcribe
and segment PDF documents into structured data.
</system_role>

<task_description>
Analyze the provided document image. Identify every layout element, its bounding box,
its category, and its textual content.
</task_description>

<categories>
Classify each element into exactly one of these categories:
section_header, text, formula, list_item, ref_item, table, image, caption,
page_header, page_footer, watermark

Rules for Categorization:
- Use "section_header" for titles and headings. Infer hierarchy based on content and
font size/boldness.
- Use "image" for charts, diagrams, or photos.
- Use "unknown" if the element is ambiguous.
</categories>

<bounding_boxes>
1. Format: [y0, x0, y1, x1] (Top-Left to Bottom-Right). You MUST provide the
coordinates in this exact order.
2. Success conditions:
- The bounding box MUST enclose the entire layout element while minimizing
unnecessary white space.
- If a character belongs to the content ALL of its pixels MUST BE CONTAINED inside
the bounding box.
3. Page Index: The current page is "page_number": {*}.
</bounding_boxes>
```

```

<extraction_rules>
- **Text Fidelity:** Extract text EXACTLY as it appears. Do NOT fix spelling or
  grammar. You MAY use any formatting that is available for a standard Markdown
  document.
- **Character Escaping:** You MUST escape any special characters that can break the
  final JSON output. Also you must escape any quotation marks.
- **Reading Order:** Sort elements by natural human reading order.
- **Special Formatting:**
  - image: Content must be an empty string "".
  - formula: Content must be LaTeX.
  - table: Content must be a Markdown table representation. TABLE CONTENT MUST NOT
    BREAK THE JSON FORMAT!
  - list_item, ref_item: Content MUST be a valid Markdown list. You MUST replace
    alternative bullet point symbols with "-". Ordered lists must start with their
    numbering followed by ".".
  - section_header: You MUST NOT use Markdown header formatting. You MUST add a "
    heading_level" field (int). Infer the level by checking the content for any
    numbering and analyzing the font size and styling of the header.
</extraction_rules>

<output_schema>
Do not return any additional text with the result.
Return a SINGLE JSON object with this exact structure:
{
  "layout_elements": [
    {
      "category": "string_(from_list)",
      "heading_level": integer (include only for headers),
      "content": "string",
      "bbox": {
        "page_number": integer,
        "box_2d": bounding_box (list[integer]) (SINGLE bounding box)
      }
    }
  ]
}
YOU MUST ENSURE THAT YOUR OUTPUT IS A VALID JSON OBJECT!
</output_schema>

```

Table A.1.: Complete list of classifications permitted to be returned by a DP implementation. Each implementation provides a mapping from their native output classifications to the standard set defined here. Some ParsingResultTypes may only be returned from a subset of these implementations.

Classification	Description
ROOT	The top-level node containing the entire document structure
TEXTS	
TITLE	The specific main title of the document
PARAGRAPH	Standard body text content
SECTION_HEADER	Section headings or subheaders within the text body
FOOTNOTE	Explanatory notes usually placed at the bottom of a page/text
LISTS	
LIST	A container node for a list of items
LIST_ITEM	An individual item within a list
REFERENCE_LIST	A container node for a list of reference items
REFERENCE_ITEM	An individual item within a reference list
FIGURES AND TABLES	
CAPTION	Descriptive text immediately accompanying a table or figure
FIGURE	Graphical elements, diagrams, or pictures
TABLE	A container node for tabular data
DOC_INDEX	A tabular node containing the TOC
TABLE_ROW	A horizontal row within a table
TABLE_CELL	An individual cell containing data within a table row
MISCELLANEOUS	
PAGE_FOOTER	Repeating page footer (page numbers, copyright, etc.)
KEY_VALUE	A specific key-value pair
PAGE_HEADER	Repeating header found at the top of pages (e.g., journal name)
KEY_VALUE_AREA	A distinct region grouped by key-value pairs (e.g., article info)
FORM_AREA	A region indicating form content (e.g., text-fields)
FORMULA	A mathematical formula
WATERMARK	A watermark from the publishing organization
FALLBACK	
UNKNOWN	Parser cannot determine the element type
MISSING	Parser returns a classification for which no mapping exists

Figure A.1.: Template of the configuration file for the content extraction evaluation using OmniDocBench. `{{OMNI_DOC_PATH}}` is to be replaced with the path to the OmniDocBench ground truth file. `{{DT_PATH}}` is to be replaced with the directory that the DP implementation’s predictions are saved in.

```
end2end_eval:
  metrics:
    text_block:
      metric:
        - Edit_dist
    display_formula:
      metric:
        - Edit_dist
    table:
      metric:
        - TEDS
        - Edit_dist
    reading_order:
      metric:
        - Edit_dist
  dataset:
    dataset_name: end2end_dataset
    ground_truth:
      data_path: {{OMNI_DOC_PATH}}
    prediction:
      data_path: {{DT_PATH}}
    match_method: quick_match
    filter:
      language: english
      data_source: academic_literature
```

List of Figures

2.1. A pair of identical bounding boxes at different IoU values. The shaded area is the intersection of the bounding boxes. As the IoU value increases, the area of the intersection approaches the area of the bounding box.	8
2.2. Naive RAG	9
3.1. Document Segmentation Pipeline	15
3.2. ParsingBoundingBox	17
3.3. ParsingResult	17
3.4. ChunkingResult	18
3.5. RichToken	23
4.1. F1 scores on the PubLayNet dataset	35
4.2. Parsing and transformation times per page	36
A.1. OmniDocBench configuration template	43

List of Tables

3.1. Abstract Functions of the Parsing Module	20
3.2. Distribution of ground truth annotations across the different element types contained in the publaynet-mini subset of the PubLayNet dataset.	27
4.1. OmniDocBench evaluation results	37
A.1. ParsingResultType	42

Acronyms

AI artificial intelligence. 1, 10, 20–22

AP average precision. 24, 25

API application programming interface. 21

AWMF Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften. 4

CNN convolutional neural networks. 10

CPG clinical practice guidelines. 1, 4, 5, 9

DLA document layout analysis. 10, 11, 16, 19–21, 24

DP document parsing. 1–3, 8–11, 13–22, 25, 27, 28, 31, 33

DTBB detected bounding box. 25

ESMO European Society for Medical Oncology. 4

FN false negative. 25

FP false positive. 25

GOT General OCR Theory. 11

GTBB ground truth bounding box. 25

IoU intersection over union. 7, 25, 26

JSON JavaScript Object Notation. 13, 14, 19, 31

LLM large language model. 1, 5–9, 12

LTRB left-top-right-bottom. 7, 15

NCCN National Comprehensive Cancer Network. 1, 4

NLP natural language processing. 1, 5, 6

OCR optical character recognition. 10, 11, 18, 20, 21

PDF portable document format. 1, 4, 8, 9, 14, 16–20, 22

PMCOA PubMed Central Open Access. 24

RAG retrieval-augmented generation. 1, 2, 7–9, 12, 15, 17, 19, 21, 22, 34

TP true positive. 25

VLM vision-language model. 6, 10–12, 20, 21

XML extensible markup language. 13, 20