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 onkologische Leitlinien

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 used generative AI, specifically Gemini 3 Pro, for the following purposes:

1. Grammar and structured writing assistance

2. Debugging and simple code generation

3. Guided generation of code documentation (e.g., python docstrings, README files)

4. Identifying literature for foundational concepts (e.g., "Which book defines the Intersection over Union?")

At no point did I use any code or text excerpts generated by AI for the purposes of my thesis without rigorously determining their correctness. Any sources cited in this thesis were extensively reviewed manually by me before including them in my research.

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

Oncology guidelines are continuously increasing in size and complexity, placing additional strain on medical personnel. Retrieval-augmented generation (RAG)-based knowledge assistants offer a promising solution for navigating these extensive documents.

The creation of such an assistant requires the preparation of the guideline documents, transforming them into a machine readable format through document parsing (DP) and chunking them into small textual units. Hereby, the characteristics of the documents and the high traceability requirements of the knowledge assistant pose significant challenges to these processes. Additionally, the fragmented methodologies and data formats of established DP implementations complicate direct comparisons.

Our research investigates the alignment of established DP datasets with oncology guidelines, identifies suitable metrics for measuring the quality of the data preparation, and proposes architectural enhancements to provide the required traceability.

We develop a modular pipeline, enabling comparisons between eight different DP implementations and four established chunking strategies. PubLayNet and OmniDocBench are identified as adequate datasets for evaluating the DP module. However, these datasets critically lack multi-page and born-digital document representations. Our study concludes that the F1 score is a suitable metric for measuring the quality of the document layout analysis (DLA), while combining the normalized edit distance and Tree-Edit-Distance-based Similarity (TEDS) is useful for evaluating the content extraction. Regarding the evaluation of the chunking module, token-wise precision and recall proved essential, while relying on combined metrics, such as the token-wise intersection over union (IoU) is discouraged.

To fulfill the requirements of the RAG-based knowledge assistant, we introduce a novel token-centric chunking methodology and improvements to the relation integration of the DP module. Ultimately, our approach facilitates visual source attribution at a high granularity but necessitates more rigorous quantitative validation in future work.

The complete source code of our experiments is available on GitHub.

Keywords: RAG, Document Segmentation, Document Parsing, Chunking, Benchmarking

Kurzfassung

Die kontinuierliche Steigerung des Umfangs und der Komplexität von onkologischen Leitlinien stellt eine zunehmende Belastung für medizinisches Personal dar. Neuartige Technologien, wie auf "retrieval-augmented generation" (RAG) basierende Wissensassistenten, bieten eine vielversprechende Lösung für die effektive Navigation dieser Leitlinien. Für die Implementierung eines solchen Assistenten müssen die Leitlinien zunächst durch "document parsing" (DP) in ein strukturiertes maschinenlesbares Format übertragen und durch "chunking" in kurze Textausschnitte aufgeteilt werden. Hierbei stellen die Eigenschaften der Leitlinien und die hohen Transparenzanforderungen des Projektes bedeutende Herausforderungen für diese Prozesse dar. Außerdem kompliziert die fragmentierte Methodik und die verschiedenen Datentypen der DP-Implementierungen einen direkten Vergleich.

Unsere Forschung untersucht die Gemeinsamkeiten zwischen bestehenden Datensätzen und den onkologischen Leitlinien, identifiziert Metriken für die Evaluation des Prozesses und erweitert diesen um den Anforderungen des Projektes gerecht zu werden.

Wir entwickeln eine modulare Dokumentsegmentierungspipeline und ermöglichen so Vergleiche zwischen acht DP-Implementierungen und vier Chunkingstrategien. PubLayNet und OmniDocBench wurden als adequate Datensätze für die Evaluierung des DP-Moduls identifiziert. Jedoch enthalten beiden Datensätzen keine mehrseitigen oder digital erstellte Dokumente. Wir schlussfolgern, dass das F1-Maß als Metrik für die Evaluation der Dokumentlayout-Analyse geeignet ist. In Kombination bieten sich die normalisierte Editierdistanz und die "Tree-Edit-Distance-based Similarity" (TEDS) für die Qualitätsanalyse der Inhaltsextraktion an. Für die Messung der Chunkingleistung befinden wir die Sensitivität und die Spezifität auf dem Token-Level als wichtigste Metriken. Von der Verwendung kombinierter Metriken, wie dem Jaccard-Koeffizienten, ist hierbei abzuraten.

Um die Anforderungen des Wissensassistenten zu erfüllen, stellen wir im Rahmen unserer Forschung eine neuartige tokenzentrierte Chunkingmethodik und verschiedene Verbesserungen für die Beziehungsintegration des DP-Moduls vor. Unsere Methodik ermöglicht eine visuelle positionsabhängige Quellenangabe mit einer hohen Granularität, erfordert allerdings zusätzliche quantitative Evaluation in zukünftigen Forschungsarbeiten.

Der vollständige, zu dieser Studie gehörende Quellcode ist auf GitHub verfügbar.

Schlüsselwörter: RAG, Dokumentensegmentierung, Document Parsing, Chunking, Benchmarking

Contents

Contents

1. Introduction

1.1. Problem Statement

Clinical practice guideliness (CPGs) are fundamental to the efficient and reliable treatment of various illnesses [1]. Oncology guidelines are a subgroup of these documents, revolving around the treatment of different forms of cancer [2]. 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 [3]. 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 [3]. 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 [3].

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 [4]. RAG is an emerging paradigm which addresses a fundamental problem of traditional large language models (LLMs) [5]. While LLMs excel at many natural language processing (NLP) tasks, they are prone to "hallucinations" and inaccurate answers when the sought information goes beyond the model's training data [6]. This poses a major obstacle for the usage of LLMs in the medical field, where accurate and reliable answers are of the highest priority [7]. RAG mitigates these drawbacks by retrieving additional context from an external knowledge source which the LLM can take advantage of during answer generation [8, 5].

The efficacy of such a RAG system is fundamentally constrained by the quality and relevancy of the context retrieved from the knowledge base [9, 10]. Therefore, the construction of the knowledge base from 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 model's 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 [11].

As LLMs are constrained by the size of their context windows, it is not feasible to store each entire guideline document as an individual entry in the knowledge base [6]. Therefore, the documents need to be split up into smaller text chunks that fit into the model's context window [6]. This process is called chunking [12].

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) [13]. 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 [5]. If the stored text chunks are too long, important information might be lost between irrelevant details [9, 10]. 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 and the chunking strategy, used to decide where to split up the oncology guidelines, need to be optimized [10].

Additionally, established implementations of popular chunking strategies do not fulfill the requirement of visual source attribution at the granularity required by the Aidvice project [14, 15, 16, 17]. 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 for measuring 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 DP benchmarks and datasets which contain documents that most closely resemble these characteristics. Through this analysis, we underline the transferability of the results achieved on these datasets to our application, while identifying unrepresented attributes which require manual comparisons. This approach allows the evaluation of various DP implementations on established datasets, without the availability of a dedicated oncology guideline dataset.

In order to evaluate the effectiveness of both DP and chunking strategies, we identify various metrics used in existing literature. We then perform a comparative analysis of the identified metrics, evaluating their applicability to 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 highly granular traceability of the chunk's content to its constituent structural elements. 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 DP techniques using the datasets and metrics identified in RQ1 and RQ2. Through this evaluation we identify promising combinations of DP 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 [18, 19]. 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 [1, 18, 19]. Instead of dictating a single definitive treatment option, CPGs focus on aiding the decision making process, promoting treatment options with proven benefits and discouraging ineffective or harmful treatments [1, 18]. As such, they aim to improve the quality of the provided health care by encouraging the translation of research into medical practice [19]. Oncology guidelines are a subgroup of this document type, focusing on the treatment and rehabilitation options for various types of cancers [19].

In order to further improve the quality and standardization of oncology guidelines [19, 3], and therefore cancer care, several prominent organizations have emerged which endorse and publish selected oncology guidelines. Prominent examples include the NCCN [2], the European Society for Medical Oncology (ESMO) [20], and, for oncology guidelines in the German language, the "Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften" (AWMF) [21]. Over the last decades, the oncology field has seen many advances in the research and treatment outcomes of many forms of cancers [3, 22]. Following these findings and advancements, the number of available treatment options has increased drastically [3]. This increase in available treatments is ultimately reflected in the increasing complexity of oncology guidelines. Kann, Johnson, Aerts, et al. [3] 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 the 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 [23]. Particularly, the guideline 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, are predominantly text-centric. 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 CPGs, 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 errors 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 Hirschberg and Manning [24], NLP "employs computational techniques for the purpose of learning, understanding, and producing human language content" (p.1). The introduction of the transformer architecture by Vaswani, Shazeer, N. Parmar, et al. [25] and the subsequent development of LLMs has revolutionized the field in recent years [26]. 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 [27]. 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 [27]. 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 [28]. During training, the tokenizer creates its dictionary by finding character pairings that occur with the highest frequency in the training data [28]. Additionally, with the multitude of different techniques for modern sub-word tokenization [29, 30, 31], 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 [32]. Every modern NLP algorithm uses embeddings as the representation of the meaning of texts [32]. 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 [32].

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

However, the dot product is not invariant to the length of the vectors $|v|$, which is defined in Equation 2.2, producing higher values for vectors of greater length [32]. This leads to skewed similarity values if vectors are not normalized prior.

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

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 [32]. 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 for NLP tasks.

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

$$a \cdot b = |a||b| \cos \theta$$
$$\frac{a \cdot b}{|a||b|} = \cos \theta \tag{2.4}$$

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 [33, 34]. 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 [33]. These models find applications involving both the comprehension and generation of multi-modal content, such as image captioning, and visual question answering [33].

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 [35]. These boxes are defined within the image's coordinate system, whose origin is typically positioned at the top-left corner of the image [36]. 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 [35]. There are multiple formats for representing these 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 [36].

2.4. Retrieval-Augmented Generation

Although LLMs have extensive general-domain knowledge due to their enormous corpora of training data, compiled from various open-domain sources [37], they struggle with tasks that require domain-specific knowledge which they did not encounter during training [6]. This can lead to "hallucinations" and inaccuracies, as the model tries to synthesize a matching answer based on its domain-wise irrelevant training data [6, 8]. RAG addresses this limitation, extending the usage of LLMs to applications requiring extensive knowledge in a specific domain [5]. 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 [5, 6].

2.4.1. Architecture of RAG Systems

While there are many advanced and extended versions of RAG systems, we will focus on the standard Naive RAG architecture for this study, as depicted in Figure 2.1 [6]. Naive RAG is based on the original RAG architecture proposed by Lewis, Perez, Piktus, et al. [5]. Naive RAG systems consist of two modules:
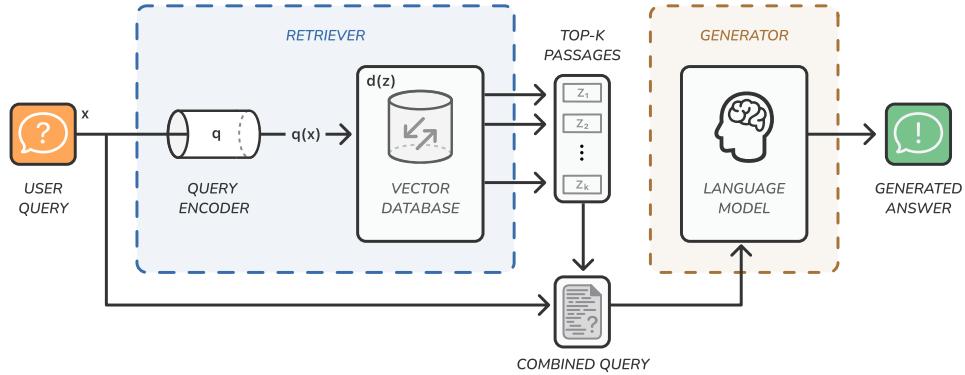
Figure 2.1.: Architecture of the Naive RAG system.

Retriever: The retriever module consists of a query encoder and an external knowledge base [5]. It is responsible for retrieving relevant context from the knowledge base, based on the user's query [6]. 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 [5, 38]. 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 [5]. To identify the relevant passages for a query x, x is first transformed into a vector embedding $q(x)$ using the retriever's query encoder q [6]. Based on the similarity scores between the query embedding and the stored chunk embeddings, the top-k documents with the highest similarity scores, are then retrieved from the database [5, 6].

Generator: The generator module is responsible for synthesizing the final answer based on the user's query and the passages retrieved from the knowledge base [5]. Firstly, the original query x and the retrieved passages are combined into a single input query [6]. The LLM is then tasked with generating the final answer y, conditioned on this combined input [6, 5].

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 [6]. Indexing begins with the preparation of the data sources into short

text passages [6]. 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 [13], and chunking, the splitting of this data into smaller text passages called chunks [6]. 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 [6]. 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 [6].

The quality of the index construction has a crucial effect on the resulting RAG system [6]. 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 [6].

2.4.3. Source Attribution

Source attribution is a mechanism that integrates transparency and traceability into the output of the RAG system by linking the generated text to its source documents [11, 39]. This allows the user to verify the LLM's claims by examining the provided sources [39]. 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 [11, 39]. While this approach enables the necessary verifiability, it places additional strain on the user, who has to find the relevant passages in the document [11]. This effect is amplified for longer documents, such as CPGs. In order to mitigate this issue, recent research has introduced the concept of visual source attribution [11]. Visual source attribution revolves around visual confirmation of the exact location of the retrieved information [11]. This is achieved by highlighting the exact region of the retrieved text inside of the document [11]. 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 [13, 40]. 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 [40, 41]. 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 be directly processed by machines [42]. 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 in.

Converting PDF documents is particularly challenging due to their variable formatting, lack of standardization, and focus on visual characteristics [42]. The format not only includes born-digital files but also 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 [40, 43]. While there are many tools and implementations available for DP [42, 41, 44, 45], 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 [46]. Additionally, by making use of lightweight models and integrating parallelization, pipeline systems can reach efficient parsing speeds [40]. While different formations are possible, most implementations consist of three different stages [13].

Document Layout Analysis (DLA): According to Q. Zhang, B. Wang, V. S.-J. Huang, et al. [13], DLA refers to the identification of the structural elements of a document, such as paragraphs, section headings, tables, figures, and interline formulas, as well as their respective bounding boxes [13, 47]. There are two types of methods for performing DLA. Uni-modal methods focus purely on visual features of the document to identify structural elements [13, 48]. Notably, convolutional neural network (CNN)- and transformer-based methods adapt models initially designed for object detection tasks, such as the YOLO [49] and DETR [50] families of models, to accurately identify structural elements in document images [13, 47]. Hereby, transformer-based methods excel at capturing global relationships between structural elements at the cost of computational intensity and expensive pre-training [13]. 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 [48, 51]. This approach allows more granular classifications and the analysis of highly complex layouts [13, 48].

Content Extraction: To extract the content of the identified structural elements different recognizers are applied to the element regions based on their classifications [13, 41, 42]. 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 [13, 52]. Popular OCR engines include EasyOCR [53] and the Tesseract OCR engine [54]. In addition to extracting content using OCR, DP implementations often provide specific recognizers for additional element types [13, 41]. Most commonly this includes a specialized model for table structure recognition, referring to the extraction of table content into structured file formats, such as the Hypertext

Markup Language (HTML), extensible markup language (XML) or Markdown [13, 42, 41, 55]. Other examples of class-specific recognizers include mathematical formula recognition and chart recognition [41, 44, 13].

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 [13, 41, 42]. 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 [16].

Implementations following a 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 [43]. Additionally, because of the sequential nature of the pipeline approach, errors from different stages propagate through the pipeline [43, 44].

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 [56]. In contrast 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 [46]. Additionally, these models demonstrate advantages in understanding the structure and hierarchy of complex documents [13]. 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 [44, 34]. However, these models are often either proprietary or require extensive computational resources [44]. Additionally, they often struggle with documents that follow more complex layouts or contain densely packed text blocks [44].

Domain-Specific VLMs: Domain-specific VLMs are trained and optimized specifically for DP [44, 13, 43]. In recent years, there has been promising developments towards domain-specific VLMs that encapsulate DLA, content extraction and relation integration into a single model [57]. 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 [57, 44]. 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 [43]. However, this approach inherently leads to the loss of information, such as the 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 [57, 43].

Recently, multiple studies have also explored multi-stage VLM-based approaches [44, 46]. These systems 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 [44]. 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 [44, 57]. Current challenges regarding the development of VLM-based approaches are the risks of "hallucinations", especially on longer documents [44, 16], as well as their high computational requirements compared to modular pipeline systems [16].

2.6. Chunking

Chunking refers to the splitting of documents into small atomic units of information called chunks [58, 6]. 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 [59, 60].

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 well with the types of modern chunking strategies. To ensure scientific stability, we therefore adopt the terminology proposed by Callan [60]. Specifically, Callan [60] 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 [60], with the advent of chunking the focus shifted towards measuring the number of tokens [58]. Modern chunking strategies have further extended this method through sliding-window approaches, which introduce a fixed overlap between neighboring chunks to preserve contextual continuity [61]. These strategies provide a simple and computationally efficient way to perform chunking [12]. However, they disregard the content of the document, which may result in chunk borders appearing inside a single word or sentence [62].

2.6.2. Semantic Passages

Semantic passages aim to enhance retrieval quality by aligning passage borders to identified subtopics of the document [62]. However, they introduce significant additional computational

complexity and may vary drastically in length [12]. 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) [63]. In recent years, there have also been novel strategies proposed for this task that leverage LLMs to determine semantically independent chunks [64].

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 [17]. While the documents in passage-based document retrieval are simple unstructured text streams [60], modern chunking techniques often process data in structured formats such as JavaScript Object Notation (JSON) or XML [16], 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 [16].

2.6.4. Metadata Attachments

In addition to their textual content, chunks can be enriched with metadata information [6]. 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 [6]. 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.
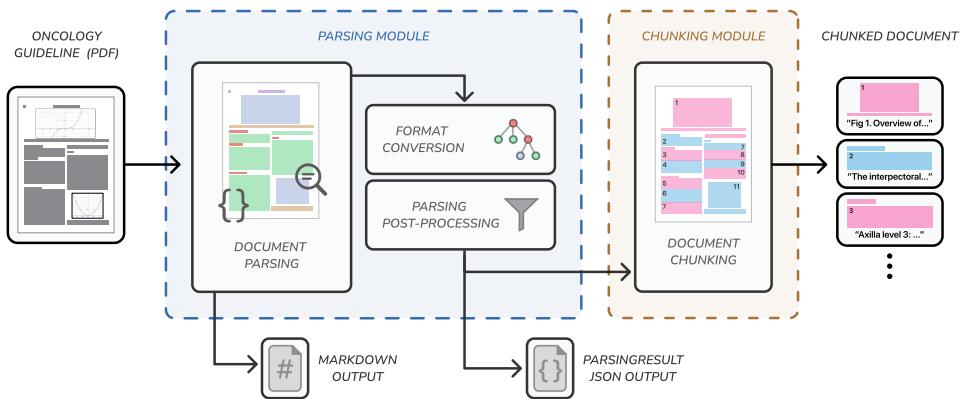
Figure 3.1.: Architecture of the document segmentation pipeline.

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 the 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 problem, 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 their 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 [45], while Docling [42] 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 implementations [41, 65], 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

```python
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.

The ParsingBoundingBox (Figure 3.2) serves as the fundamental data type for denoting 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 [16] and Google Document AI LayoutParser [66], 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 document elements through parent-child relationships, ensuring a lossless representation of the original document. The ParsingResult data type (Figure 3.3) represents a node inside of this tree structure.

```python
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 extracted 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 information 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, and the metadata from the root node of the preceding ParsingResult tree.

```
                                    class Chunk:
    class ChunkingResult:               id: str
        chunks: list[Chunk]             content: str
        metadata: dict                  metadata: dict
                (a)                     geom: list[ParsingBoundingBox]
                                                (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. The significant variability between the data structures of the various DP systems

makes this the most complex task of the module, with its inner workings and required steps being highly dependent on the specific data types. Typical steps include the transformation of the element's bounding box coordinates into the normalized coordinates of the Parsing-BoundingBox, the normalization of the elements classification into ParsingResultType, and the modeling of recognized hierarchical relationships in the ParsingResult tree. The abstract `_transform` function encapsulates 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 [41], 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 [42], 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 [16], 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 [67], 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 for 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
`_parse`	PDF document	Custom data format
`_transform`	Custom data format	ParsingResult
`_get_md`	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 3.1 need to be implemented.

3.3.1. Unstructured.io

Unstructured.io is a prominent provider for DP, offering both a cloud-based API and an open-source library. For our study, we will focus on the open-source library version of Unstructured.io [45, 55]. While the developers themselves explicitly highlight that the open-source library is not suited for large-scale production environments [55], its inclusion within the documentation of popular RAG frameworks, such as Langchain [14] and LlamaIndex [15] 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, an uni-modal vision transformer, to perform DLA [55, 68]. The library also includes a specialized model for table structure recognition [55].

3.3.2. Docling

Docling, which was developed by IBM in 2022, is one of the most popular available open-source DP libraries [42, 16]. 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 [42]. Docling offers two different approaches for DP:

Parsing pipeline: Docling's processing pipeline consists of three components: a PDF back-end called DoclingParse, an internal model pipeline containing multiple AI models, and a post-processing stage [16, 42]. 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 [69] as well as for DLA with their Heron model [47]. Heron is derived from RT-DETR [70], an uni-modal vision transformer, and retrained on DocLayNet [71], Docling's own dataset for DLA. During DLA, identified bounding boxes are compared and intersected with bounding boxes retrieved from the PDF backend to provide more accurate localizations [16]. 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 finally combined into the DoclingDocument data type [16].

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 [72]. The model is very compact, consisting of around 258 million parameters [73, 72]. 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 [73].

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 [41, 44, 65].

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 [41, 67]. This metadata includes the language of the document, the document's page dimensions and the identification of scanned documents [41]. The pipeline then uses models from the DP model library PDF-Extract-Kit for DLA and content extraction [41, 65, 74, 75]. The model used for DLA is a fine-tuned version of LayoutLMv3, a multi-modal model [41, 51]. For content extraction, special models for formula and table structure recognition are employed by the pipeline.

During the final post-processing stage, overlapping elements and unneeded elements are removed and the reading order of the document elements is inferred using a segmentation algorithm [41]. 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 MinerU 2.5, 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 [44]. 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 [44].

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 [76]. While Google offers a more capable model in the form of Gemini 2.5 Pro, we follow the sentiment from Niu, Z. Liu, Gu, et al. [44], 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 [76], 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 [76, 77]. We follow the documentation provided by Google on harnessing Gemini's image understanding capabilities [77] 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 creators of LlamaIndex, a popular framework for building RAG systems and workflows [15, 78]. While there is no official information about the architecture used for the DP system behind LlamaParse, its marketing as a "GenAI-native document parser" [78] 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 [66]. In contrast to other services such as Google Document AI Enterprise Document OCR [79], 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 downstream RAG applications.

Prominent implementations of chunking strategies, such as the ones found in the RAG frameworks LlamaIndex [15] and Langchain [14], 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 HierarchicalChunker [42, 16], 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.

```python
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 [80] 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 [11], 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 [61]. 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 [12, 62]. 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 [62, 17, 14].

The delimiters used in this implementation are adapted from LangChain's RecursiveCharacterTextSplitter [14, 17], with punctuation added to better identify sentence endings, as suggested by B. Smith and Troynikov [58]. This results in the following delimiters: `["\n\n"`, `"\n"`, `"."`, `"!"`, `"?"`, `"␣"`, `""]`.

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 [62]. 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 [12]. 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) [81, 82]. 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 cosine distance, which is the inverse of the sentences similarity, denotes a topical shift between these sentences [17].

We then insert a breakpoint between sentences which have a distance that is higher 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 HybridChunker from the Docling toolkit [16], 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 headings 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_h is imposed on the length S of the section heading tokens. If adding
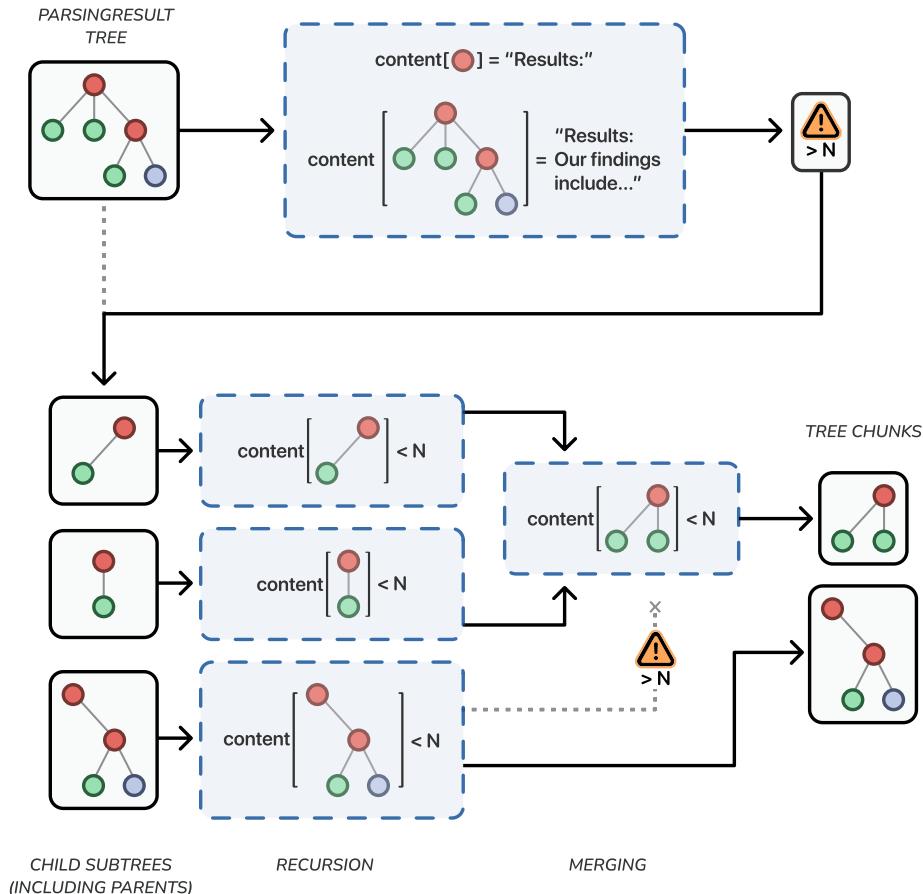
Figure 3.6.: Operational logic of the hierarchical chunker. The initial ParsingResult tree exceeds the maximum token limit and is decomposed into its child subtrees. Each child subtree contains the root node, retaining the context from the document hierarchy. In this example, the subtrees do not exceed the maximum chunk length and are not split further. The algorithm then merges adjacent child subtrees if their combined tree satisfies the size constraint.

an additional heading causes S to exceed B_h, 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 target 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 chunking 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 [83]. Although there are multiple datasets available for this task [84, 71, 40], we will use the PubLayNet dataset [85] 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 [40, 85]. This format closely resembles the format of the oncology guideline documents which makes it a suitable choice for this evaluation. Comprised of over 360.000 automatically annotated document pages collected from PubMed Central Open Access (PMCOA), PubLayNet is one of the largest datasets for DLA [85]. 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 [86]. As seen in Table 3.2, the subset contains around 5000 ground truth annotations for elements from 5 different classes.

To assess the performance of different predictors on object detection tasks such as DLA, average precision (AP) is the most commonly used metric [87]. Previous evaluations of DLA models on the PubLayNet dataset also use a version of this metric [88]. However, AP relies on the predictor's confidence values, indicating how confident the predictor is about a predicted bounding box and class label. As some of the DP implementations provide

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. Element types are disproportionally represented, with "Text" making up a large majority of the dataset.

"hard-predictions", which do not contain any confidence values, the AP is not a viable metric for the purposes of this study [89, 90].

For this reason, we will compare the implementations based on their achieved F1 scores. Similarly to AP, this metric takes into account two important measures for object detectors: precision and recall [91]. According to Padilla, Passos, Dias, et al. [91], "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, the detected bounding boxes (DTBBs) are first classified into true positives (TPs) and false positives (FPs).

The classification relies on how accurately the DTBBs align with the ground truth bounding boxes (GTBBs), which is measured through the intersection over union (IoU). According to the definition from Kaur and Singh [92], the IoU between two bounding boxes BB_a and BB_b is defined as described in Equation 3.1. 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.

$$IoU(BB_a, BB_b) = \frac{\text{Area of intersection of } BB_a \text{ and } BB_b}{\text{Area of union of } BB_a \text{ and } BB_b} \tag{3.1}$$

A DTBB is classified as a TP if there exists a GTBB from the same class such 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 criteria, 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 Lipton, Elkan, and Narayanaswamy [93], for a class i, precision and recall are formulated as described in Equation 3.2 and Equation 3.3. Hereby tp_i, fp_i, and fn_i denote the count of their respective classification for bounding boxes of class i.

$$\text{Pr}_i = \frac{tp_i}{tp_i + fp_i} = \frac{tp_i}{\text{all detections}_i} \tag{3.2}$$

$$\text{Re}_i = \frac{tp_i}{tp_i + fn_i} = \frac{tp_i}{\text{all ground truths}_i} \tag{3.3}$$

(a) $IoU = 0.5$ (b) $IoU = 0.75$ (c) $IoU = 0.95$
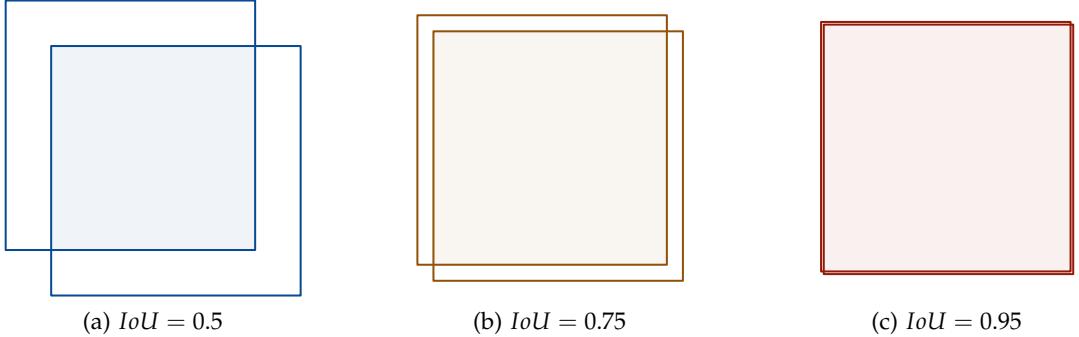
Figure 3.7.: A pair of identical bounding boxes at different IoU values. The shaded area is the intersection of the bounding boxes. As the IoU value increases, the area of the intersection approaches the area of the bounding box.

The F1 score is the weighted harmonic mean between precision and recall and is calculated as defined in Equation 3.4 [91]. The metric is calculated for a single class i 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 [91]. A F1 score calculated at an IoU threshold T% is commonly referred to as F1@T [94].

$$\text{F1}_i = \frac{2 \cdot \text{Pr}_i \cdot \text{Re}_i}{\text{Pr}_i + \text{Re}_i} \tag{3.4}$$

For scenarios with multiple classes, such as the PubLayNet dataset, the weighted F1 score can be used to assess the overall performance of the predictor. The weighted F1 score is the weighted average of the single-class F1 scores [95]. For a dataset with C different classes, the calculation of the weighted F1 score is described in Equation 3.5. Hereby, w_i denotes the proportion of GTBBs that belong to class i [95].

$$\text{F1}_{\text{weighted}} = \sum_{i=1}^{C} w_i \cdot \text{F1}_i \tag{3.5}$$

The DP implementations will be evaluated on both their single-class and weighted F1 scores. Specifically, their (weighted) 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 [96]. This rewards implementations that provide more accurate bounding boxes [96]. F1@50 is chosen, as a threshold of 50 percent is one of the most commonly used threshold values for metrics in object detection [91]. To calculate these metrics, the `faster-coco-eval` package is used to determine the recall and precision values at the IoU thresholds [97].

3.5.2. Content Extraction Evaluation

The goal of the 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 [40, 98, 44, 46, 57, 16], 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 [40]. In total, the benchmark contains 1355 single-page PDF documents [40]. 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 [51] for DLA and PaddleOCR [98], UniMERNet [75], and GPT-4o [99] 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' structural elements are extracted from the Markdown texts. Since the Markdown format is purely textual, this step is primarily performed using regular expression matching [40]. The format of tables varies depending on the DP implementation, with Markdown, HTML, and LaTeX formats being possible. To prevent interferences 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 [40].

Matching: The extracted elements are now matched to ground truth elements of the same type through a process called Adjacency Search Match [40]. 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 [40]. 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 [40].

Metric calculation: OmniDocBench provides a multitude of different evaluation metrics for the extracted elements [40]. 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 [100, 101]. $\lambda \notin \Sigma$ is the empty string. The metric denotes the minimum cost of transforming X into Y through weighted elementary edit operations. According to Yujian and Bo [101], an "elementary edit operation $[T]$ is a pair $(a, b) \neq (\lambda, \lambda)$, often written as $a \to b$, where both a and b are strings of lengths 0 or 1." (p.1) There are three elementary edit operations: insertions ($\lambda \to a$), substitutions ($a \to b$), and deletions ($b \to \lambda$). 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 3.6 [101].

$$\gamma(T_{X,Y}) = \sum_{i=1}^{l} \gamma(T_i) \tag{3.6}$$

Following the definitions from Yujian and Bo [101], given $X, Y \in \Sigma^*$, the GLD is then defined as in Equation 3.7.

$$\text{GLD}(X, Y) = \min\{\gamma(T_{X,Y})\} \tag{3.7}$$

The GLD is not normalized with respect to the lengths of X and Y. Therefore, transforming short strings with the same amount of edit operations as long strings is not additionally penalized, even though the number of edit operations represents a larger proportion of the total content. The normalized GLD, which is defined in Equation 3.8, 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 $[0, 1]$ [101].

$$\alpha = \max\{\gamma(a \to \lambda), \gamma(\lambda \to b), a, b \in \Sigma\}$$
$$d_{\text{N-GLD}}(X, Y) = \frac{2 \cdot \text{GLD}(X, Y)}{\alpha \cdot (|X| + |Y|) + \text{GLD}(X, Y)} \tag{3.8}$$

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 DP implementations is a key requirement for our evaluation. TEDS measures the similarity between two HTML tables. Therefore, in order to calculate this metric, extracted LaTeX tables are first converted into HTML format [40]. TEDS builds on top of the tree edit distance proposed by Pawlik and Augsten [102]. 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) [102].

   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 [103].

   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 depending on the type and attributes of the node. If one of the nodes 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. Otherwise, if the leaf nodes differ in their content, their substitution cost is the normalized GLD between their contents [103].

   Following the definition from Zhong, ShafieiBavani, and Yepes [103], the TEDS between two table trees *F* and *G* is then computed as in Equation 3.9. 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 $[0, 1]$ [103].

$$\text{TEDS}(F, G) = 1 - \frac{\text{EditDist}(F, G)}{\max(|F|, |G|)} \tag{3.9}$$

   In addition to the regular TEDS, OmniDocBench provides a structure-only version of this metric [40]. 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 single and double-column layouts of the oncology guidelines. The normalized edit distance of the reading order is used for this evaluation [40]. 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 [40]. 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 I' denotes the ordering returned by the DP implementation. The normalized GLD 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

Improvements in parsing quality are rendered inconsequential 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 [58]. Established RAG evaluation frameworks, such as Ragas [104], evaluate the quality of the retrieved context by using a LLM to decide whether a retrieved chunk is relevant to the query [104]. While this approach is a sensible and common choice for evaluating the performance of the retriever, it is not optimal for our evaluation, as it fails to measure the ratio of relevant information inside the retrieved chunks.

B. Smith and Troynikov [58] 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 computational strain and distractions during generation [105, 58]. 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 [58]. 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 still in development and may not be made available to the general public, we were permitted to use both the documents and the QA pairs in their current state for our evaluation.

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: B. Smith and Troynikov [58] 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 $c = \{c_1, c_2, \ldots, c_l\}, c_i \in C$ 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 [106]. 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, B. Smith and Troynikov [58] account for this redundancy by noting that the cardinality of T_r includes the multiplicity of included tokens [58]. To formulate this logic with the needed mathematical rigor, we define M_r as the multiset over T_r. $M_r : T_r \to \mathbb{N}_0$ is a function such that if $M_r(t) = k > 0, t \in T_r$, then t appears with multiplicity k in M_r [106]. The cardinality of the multiset is defined as $|M_r| = \sum_{t \in T_r} M_r(t)$ [106]. 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.

Q: Which therapy is recommended for a patient with condition B?

T_r (M_r)	T_e	$T_r \cap T_e$
Therapy 1 is recommended for patients with condition A. Therapy 2 is recommended for patients with condition B, which includes multiple rounds of medication X.	Therapy 1 is recommended for patients with condition A. Therapy 2 is recommended for patients with condition B, which includes multiple rounds of medication X.	Therapy 1 is recommended for patients with condition A. Therapy 2 is recommended for patients with condition B, which includes multiple rounds of medication X.
RETRIEVED CHUNK	*GROUND TRUTH*	*RETRIEVED GROUND TRUTH*

Figure 3.8.: Visual example for the methodology of the chunking evaluation. For a given query the tokens from the blue area are retrieved from the chunk corpus, while the tokens from the yellow area are needed to answer the question. The set of correctly retrieved tokens is denoted by the red area.

Following this principle, B. Smith and Troynikov [58] propose three distinct evaluation metrics. For each evaluated chunking strategy, the mean of each metric over the QA pairs is

reported [58].

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 3.10 [58].

$$\text{Pr}_q(C) = \frac{|T_e \cap T_r|}{|M_r|} \tag{3.10}$$

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 3.11 [58].

$$\text{Re}_q(C) = \frac{|T_e \cap T_r|}{|T_e|} \tag{3.11}$$

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 3.12 [58].

$$\text{IoU}_q(C) = \frac{|T_e \cap T_r|}{|T_e| + |M_r| - |T_e \cap T_r|} \tag{3.12}$$

4. **Precision$_\Omega$:** Precision$_\Omega$ 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 [58].

Figure 3.9.: Visual example for the calculation of the chunking evaluation metrics. Calculating the token-wise IoU is analog to the calculation of the IoU of two bounding boxes.

Deciding on the specific configurations for evaluating the chunking strategies is not trivial, as there is no universally accepted "best" set of parameters. The optimal values are highly-dependent on the information density and attributes of the document corpus [107].

Furthermore, there is limited scientific research that quantifies the effect of specific parameter configurations across multiple strategies.

To address this uncertainty, we evaluate each chunking strategy across a range of parameters to identify optimal settings for our application while investigating how different parameters influence retrieval performance. To ensure a thorough investigation while keeping the number of configurations manageable, we identify a range of options for each parameter based on scientific literature and recommendations from established RAG frameworks.

1. **Maximum chunk size (N):** The maximum chunk size is the most critical tuneable chunking parameter, as it directly impacts the granularity of the retrieved context. Bhat, Rudat, Spiekermann, and Flores-Herr [107] suggest that smaller chunk sizes ($N \leq 128$) are effective for datasets that require concise fact-based answers, such as biomedical documents, whereas larger chunks ($512 \leq N \leq 1024$) are necessary for documents where information is spread out across long paragraphs. LlamaIndex recommends a maximum token length of 1024 tokens for most applications, finding that further increases lead to decreased performance [108]. In order to facilitate an extensive review of different chunk sizes, we evaluate every strategy at $N \in \{128, 256, 512, 1024\}$.

2. **Chunk overlap (O):** While sliding-window approaches are common in practice, empirical research on the specific effects of overlap remains sparse. Configurations used in current literature vary drastically, from the variable $0.5 \cdot N$ experimented with by B. Smith and Troynikov [58] to the fixed smaller values used by both R. Qu, Tu, and Bao [12] and LlamaIndex, with the latter proposing a very small default overlap of 20 tokens. In order to strike a balance between these configurations and test the impact of including an overlap, we evaluate the strategies at $O \in \{0, 0.20 \cdot N\}$.

3. **Breakpoint percentile (Q):** This parameter determines the sensitivity of the breakpoint-based semantic chunker to topical shifts in the document. Setting a larger Q hereby results in the chunker being more selective, resulting in fewer, larger chunks with more pronounced topical shifts. We experiment with $Q \in \{0.7, 0.9\}$.

4. **Minimum chunk size (M):** To ensure that semantic chunks scale with the maximum chunk size, the minimum chunk size needs to scale accordingly. We set $M = 0.5 \cdot N$.

5. **Heading token budget (B_h):** In hierarchical chunking, the heading budget prevents section headings from taking up a majority of the chunk's content. For our evaluation, we set $B_h = 0.4 \cdot N$.

In total, we evaluate 28 different chunking configurations across four different strategies.

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. To optimize the performance of this hardware, especially for the locally deployed VLM-based DP approaches, we utilize Apple's MLX engine [109] when possible.

4. Results

4.1. Document Parsing Evaluation

Building on the benchmarks and metrics introduced in Section 3.5, we present the quantitative evaluation of the selected DP implementations. This analysis covers the performance results for both the DLA and content extraction evaluations, followed by a comparison of the implementations' processing speeds based on their average parsing time per page.

4.1.1. Document Layout Analysis Evaluation

The results of the evaluation of each DP implementations' DLA performance on the Pub-LayNet dataset are summarized in Table 4.1.

Regarding the F1@50 scores (Table 4.1a), MinerU 2.5 VLM achieved the highest overall weighted score of 0.8599, followed by Docling (0.8555) and MinerU 2.5 Pipeline (0.8545).

In terms of single-class F1@50 performance, Document AI recorded the highest values in three out of the five categories, namely "title" (0.9789), "table" (0.9911), and "figure" (0.9848). However, both Document AI and LlamaParse recorded a score of 0.0000 for the "list" category, with LlamaParse additionally reporting the same score for the "figure" type. Across all implementations, the lowest scores were also consistently observed in either the "list" or "figure" category. Hereby, "list" recorded the lowest maximum F1 score among all element types, with Docling reaching the highest value of 0.8022.

For the "text" category, which represents more than 72 percent of the annotations in the `publaynet-mini` dataset (Table 3.2), MinerU 2.5 VLM (0.9119) achieved the highest score, with MinerU2.5 Pipeline (0.8735) following in second place. While the performance of the two MinerU systems remained similar across most element types, they diverged by a margin of 0.4026 for the "figure" category, with the pipeline approach (0.6534) scoring higher than the VLM-based approach (0.2508). Granite Docling (0.5989) and LlamaParse (0.7031) recorded the lowest weighted F1 scores, with one of these two implementations consistently ranking last in every individual category.

Pivoting to the F1@50:95 scores (Table 4.1b) and the corresponding decrease in reported values (Table 4.2), results in a shift in the model rankings. Increasing the IoU requirements sees Docling claim the highest overall weighted F1 score with 0.7304 as MinerU 2.5 Pipeline (0.7189) takes the second rank. Meanwhile, its VLM-based counterpart (0.6568) drops to fifth place behind Unstructured.io (0.6691) and Document AI (0.6663). This trend is also reflected in the "text" category where Docling (0.8206, $\Delta = -0.0481$) claims the top spot in front of MinerU 2.5 Pipeline (0.8097, $\Delta = -0.0638$) and VLM (0.8032, $\Delta = -0.1087$).

	text	title	list	table	figure	all
Docling	0.8687	0.8775	**0.8022**	0.9530	0.5951	<u>0.8555</u>
Document AI	0.7830	**0.9789**	0.0000	**0.9911**	**0.9848**	0.8197
Gemini 2.5 Flash	0.8242	0.7619	0.4107	0.8725	0.6530	0.7923
Granite Docling	0.6737	0.6309	0.6329	0.9001	0.1811	0.5989
LlamaParse	0.7711	0.6370	0.0000	0.6831	0.0000	0.7031
MinerU 2.5 Pipeline	<u>0.8735</u>	<u>0.9558</u>	0.4771	0.9784	<u>0.6534</u>	0.8545
MinerU 2.5 VLM	**0.9119**	0.8822	0.5702	<u>0.9796</u>	0.2508	**0.8599**
Unstructured.io	0.8123	0.8032	<u>0.6638</u>	0.9337	0.6138	0.7780

(a) F1@50

	text	title	list	table	figure	all
Docling	**0.8206**	<u>0.6311</u>	**0.7406**	0.9143	0.4952	**0.7304**
Document AI	0.7136	**0.6595**	0.0000	**0.9669**	**0.9524**	0.6663
Gemini 2.5 Flash	0.7347	0.5002	0.3195	0.7636	<u>0.5822</u>	0.6441
Granite Docling	0.6241	0.4302	0.5694	0.8497	0.1608	0.5141
LlamaParse	0.7240	0.3057	0.0000	0.6594	0.0000	0.5738
MinerU 2.5 Pipeline	<u>0.8097</u>	0.6170	0.4331	0.9407	0.5436	<u>0.7189</u>
MinerU 2.5 VLM	0.8032	0.4461	0.4906	<u>0.9409</u>	0.2034	0.6568
Unstructured.io	0.7583	0.6029	<u>0.5722</u>	0.8707	0.4987	0.6691

(b) F1@50:95

Table 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 weighted F1 score. Highest values are bolded and second highest values are underlined. Higher values are preferred.

	text	title	list	table	figure	all
Docling	<u>0.0481</u>	0.2464	<u>0.0616</u>	0.0387	0.0999	0.1251
Document AI	0.0695	0.3194	-	<u>0.0242</u>	<u>0.0324</u>	0.1535
Gemini 2.5 Flash	0.0894	0.2617	0.0911	0.1088	0.0708	0.1482
Granite Docling	0.0496	<u>0.2007</u>	0.0635	0.0504	**0.0203**	**0.0848**
LlamaParse	**0.0471**	0.3314	-	**0.0237**	-	0.1293
MinerU 2.5 Pipeline	0.0638	0.3388	**0.0440**	0.0377	0.1098	0.1356
MinerU 2.5 VLM	0.1087	0.4361	0.0796	0.0387	0.0474	0.2031
Unstructured.io	0.0540	**0.2003**	0.0916	0.0630	0.1151	<u>0.1088</u>

Table 4.2.: Decrease in reported scores ($-\Delta$) when transitioning from F1@50 to F1@50:95 scores on the PubLayNet dataset.

The "title" element type sees the most substantial decrease in reported scores across the individual categories, ranging from Unstructured.io's 0.2003 up to the 0.4361 decrease reported by MinerU 2.5 VLM. The smallest average decrease was observed for the "table" elements ($\bar{\Delta} = 0.0482$). Granite Docling ($\Delta = -0.0848$) and Unstructured.io ($\Delta = -0.1088$) report the smallest change in overall weighted F1 scores.

4.1.2. Content Extraction Evaluation

Table 4.3 presents the results of the content extraction evaluation using OmniDocBench. The evaluation was performed for the subset of English scientific literature documents, yielding results for a total of 129 single-page PDF documents. MinerU 2.5 VLM recorded the highest overall score of 94.3450 as Miner U 2.5 Pipeline (90.6538) followed in second place.

	$\text{Text}^{Edit} \downarrow$	$\text{Table}^{TEDS} \uparrow$	$\text{Table}^{S-TEDS} \uparrow$	$\text{Read}^{Edit} \downarrow$	Overall $\uparrow$
Docling	0.0780	66.3294	85.2592	0.0791	83.5388
Document AI	0.0452	62.9350	74.8045	<u>0.0261</u>	85.2689
Gemini 2.5 Flash	0.0455	65.3683	72.6077	0.0409	85.5765
Granite Docling	0.1365	64.2110	70.0415	0.0890	80.5537
LlamaParse	0.1688	61.0700	74.7680	0.1905	75.0460
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**	**94.3450**
Unstructured.io	0.0836	64.3603	82.0875	0.1161	81.4640

Table 4.3.: Results of the content extraction evaluation on the OmniDocBench benchmark. $S - TEDS$ is the structure-only TEDS. *Edit* is the normalized edit distance. Overall is the mean of $1 - \text{Text}^{Edit}$, Table^{TEDS}, and $1 - \text{Read}^{Edit}$. For metrics marked with $\uparrow$ higher values are preferred, while $\downarrow$ denotes that lower values are better. Best values are bolded and second-best values are underlined.

MinerU 2.5 VLM also recorded the highest value for each of the individual metrics. For the normalized edit distance, MinerU 2.5 VLM (0.0247) led by a significant margin in front of MinerU 2.5 Pipeline (0.0439), Document AI (0.0452), and Gemini 2.5 Flash (0.0455). For TEDS, the MinerU 2.5 implementations were the only ones of the evaluated approaches that reached a score higher than 80 (VLM: 86.0939, Pipeline: 80.2155), with Docling (66.3294) being the closest competitor. The same pattern is observed for the structure-only variant of the metric, as MinerU 2.5 VLM (92.8198) and Pipeline (90.1323) scored highest and Docling (85.2592) came third.

The reading order edit distance is the only metric where the second rank was not occupied by MinerU 2.5 Pipeline (0.0386). Instead, Document AI (0.0261) took its place. MinerU 2.5 VLM (0.0059) reached the first place with a normalized edit distance that is less than one fourth of the score achieved by Document AI.

The lowest values for both the overall and individual metric scores were consistently reported by either LlamaParse or Granite Docling. LlamaParse (75.0460) is the only implemen-

tation that scored less than 80 in the overall metric, with Granite Docling (80.5537) scoring slightly above.

4.1.3. Processing Times

We evaluated the average processing time per page for each of the different DP implementations across the PDF documents used for the DLA and content extraction evaluation. In total, the mean parsing time was evaluated across 629 document pages. The observed mean processing time per page in seconds and standard deviation for each implementation are displayed in Table 4.4.
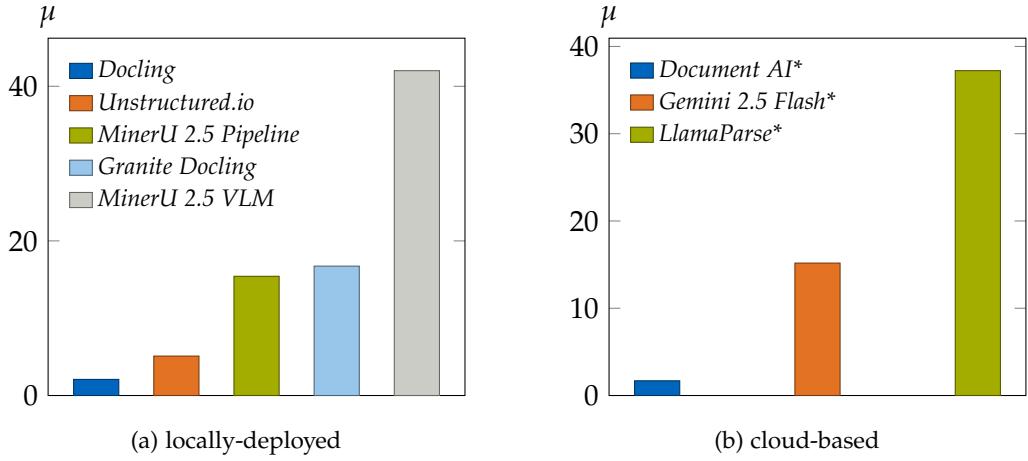
Figure 4.1.: Mean parsing times for locally-deployed (a) and cloud-based (b) DP implementations. Values are reported in seconds per page.

The evaluated DP implementations varied greatly in their mean processing time per page. Overall, Document AI achieved the shortest parsing time per page of 1.7 seconds. Only considering implementations that were evaluated on local hardware sees Docling claim the top spot with an average of 2.1 seconds per page. Except for Unstructured.io ($\mu = 5.1$), all other methods reported parsing times above 15 seconds per page. Hereby, MinerU 2.5 VLM reported the slowest processing speed with an average of 42 seconds per page. Among cloud-based services, LlamaParse had the slowest parsing speed, processing each page for an average of 37.2 seconds.

Document AI ($\sigma = 0.4101$) and Docling ($\sigma = 1.3166$) also reported the smallest standard deviation. Despite having a faster mean parsing time than both LlamaParse and MinerU 2.5 VLM, Granite Docling ($\mu = 16.74$, $\sigma = 63.5661$) reported the highest standard deviation among all evaluated methods.

	seconds per page	
	μ	σ
Docling	<u>2.0839</u>	<u>1.3215</u>
Document AI*	**1.6863**	**0.4101**
Gemini 2.5 Flash*	15.1748	13.1830
Granite Docling	16.7379	63.5661
LlamaParse*	37.2230	37.8810
MinerU 2.5 Pipeline	15.4215	20.2052
MinerU 2.5 VLM	42.0182	35.4030
Unstructured.io	5.1010	5.5998

Table 4.4.: Mean (μ) and standard deviation (σ) of the DP implementations' parsing times per page. Times are reported in seconds. Fastest times are bolded and second fastest times are underlined. Lower values are preferred. Cloud-based services are marked with "*".

4.2. Chunking Evaluation

We present the evaluation results for 28 different chunking configurations on both the manually-annotated QA dataset from the TUM university hospital as well as the datasets created for the evaluation performed by B. Smith and Troynikov [58]. Each document corpus was converted into a chunk corpus for each of the chunking configurations. Embeddings were created using OpenAI's `text-embedding-3-small` model [110] and Chroma [111] was used for creating and retrieving from the vector database. During retrieval, the number of retrieved passages was determined dynamically as the number of chunks that contain tokens from the references included in each QA pair.

4.2.1. Oncology Guideline QA Dataset

Table 4.5 presents the results of the chunking evaluation on the expert-annotated QA pairs. In total, the dataset contains 46 QA pairs, created over a document corpus of eleven German oncology guidelines published by AWMF. Each document was parsed using Docling before chunk creation. Our results indicate that the impact of specific parameters varies across different chunking strategies.

1. **Fixed-size chunking:** This strategy reached the highest token-wise recall of all tested configurations at a maximum chunk size $N = 512$ with an overlap between chunks of $O = 104$ ($0.2 \cdot N$). Recall improved consistently as with increasing chunk size up to this point, after which it began to decrease. Introducing overlap yielded higher recall than the non-overlapping configuration with the same maximum chunk length while maintaining steady precision. An exception to this trend was observed at $N = 1024$, where overlap added no benefit to recall. Notably, the configurations $N = 256$ ($O = 52$)

		N	IoU	Recall	Precision	Precision$_\Omega$
Fixed-Size	$O = 0$	128	0.10 ± 0.11	0.29 ± 0.29	0.12 ± 0.12	0.42 ± 0.11
	$O = 26$	128	0.11 ± 0.10	0.35 ± 0.29	0.12 ± 0.11	0.38 ± 0.12
	$O = 0$	256	0.09 ± 0.10	0.33 ± 0.31	0.10 ± 0.11	0.30 ± 0.12
	$O = 52$	256	0.09 ± 0.07	0.43 ± 0.33	0.10 ± 0.08	0.27 ± 0.10
	$O = 0$	512	0.09 ± 0.07	0.50 ± 0.32	0.10 ± 0.08	0.19 ± 0.08
	$O = 104$	512	0.08 ± 0.06	$\mathbf{0.56 \pm 0.34}$	0.09 ± 0.07	0.17 ± 0.07
	$O = 0$	1024	0.07 ± 0.07	0.49 ± 0.36	0.07 ± 0.07	0.12 ± 0.06
	$O = 208$	1024	0.06 ± 0.05	0.49 ± 0.37	0.06 ± 0.05	0.12 ± 0.06
Recursive	$O = 0$	128	0.12 ± 0.15	0.26 ± 0.28	0.16 ± 0.17	$\underline{0.59 \pm 0.15}$
	$O = 26$	128	0.10 ± 0.11	0.30 ± 0.26	0.13 ± 0.11	0.45 ± 0.12
	$O = 0$	256	0.12 ± 0.13	0.36 ± 0.31	0.14 ± 0.14	0.39 ± 0.14
	$O = 52$	256	0.10 ± 0.08	0.40 ± 0.29	0.11 ± 0.09	0.30 ± 0.11
	$O = 0$	512	0.12 ± 0.13	0.50 ± 0.37	0.13 ± 0.13	0.25 ± 0.12
	$O = 104$	512	0.09 ± 0.07	0.51 ± 0.32	0.10 ± 0.08	0.19 ± 0.08
	$O = 0$	1024	0.09 ± 0.09	0.52 ± 0.39	0.09 ± 0.09	0.14 ± 0.07
	$O = 208$	1024	0.06 ± 0.05	$\underline{0.55 \pm 0.36}$	0.06 ± 0.05	0.12 ± 0.06
Semantic	$Q = 70$	128	$\underline{0.13 \pm 0.16}$	0.29 ± 0.29	$\underline{0.17 \pm 0.18}$	0.58 ± 0.14
	$Q = 90$	128	0.12 ± 0.15	0.27 ± 0.29	0.16 ± 0.17	$\underline{0.59 \pm 0.15}$
	$Q = 70$	256	0.12 ± 0.12	0.36 ± 0.31	0.14 ± 0.13	0.37 ± 0.13
	$Q = 90$	256	0.11 ± 0.11	0.33 ± 0.30	0.13 ± 0.13	0.40 ± 0.13
	$Q = 70$	512	0.10 ± 0.11	0.44 ± 0.33	0.11 ± 0.11	0.24 ± 0.11
	$Q = 90$	512	0.10 ± 0.10	0.46 ± 0.35	0.11 ± 0.10	0.24 ± 0.11
	$Q = 70$	1024	0.07 ± 0.07	0.48 ± 0.37	0.07 ± 0.07	0.13 ± 0.06
	$Q = 90$	1024	0.08 ± 0.06	0.52 ± 0.36	0.08 ± 0.07	0.14 ± 0.07
Hierarchical	$B_h = 51$	128	$\mathbf{0.17 \pm 0.20}$	0.32 ± 0.28	$\mathbf{0.22 \pm 0.22}$	$\mathbf{0.62 \pm 0.17}$
	$B_h = 102$	256	0.12 ± 0.13	0.34 ± 0.30	0.15 ± 0.14	0.38 ± 0.20
	$B_h = 204$	512	0.10 ± 0.12	0.34 ± 0.33	0.11 ± 0.12	0.24 ± 0.16
	$B_h = 409$	1024	0.09 ± 0.11	0.33 ± 0.34	0.09 ± 0.12	0.13 ± 0.11

Table 4.5.: Results of the chunking evaluation on the manually annotated medical QA pairs. All values are reported as $\mu \pm \sigma$. Highest μ are bolded and second-highest are underlined.

and $N = 512$ ($O = 0$) reported identical IoU and precision values, despite the latter achieving a higher recall ($\Delta_{Re} = 0.07$). Precision displayed a linear negative relationship with the maximum chunk size, decreasing as N increased.

2. **Recursive character chunking:** For recursive character chunking, recall increased monotonically as chunk size increased, reporting its highest value of 0.55 for the largest maximum chunk size $N = 1024$. While overlap positively influenced the recall value, the effect was of a smaller magnitude than in fixed-size chunking. Hereby, the configurations for $N = 512$ showed the smallest positive effect upon the introduction of an overlap value ($\Delta_{Re} = 0.01$). Additionally, the inclusion of overlap had a negative effect on the reported precision, leading to an average decrease of $\bar{\Delta}_{Pr} = 0.03$. Precision$_\Omega$ values were consistently higher than those of fixed-size configurations, although this gap diminished with increasing maximum chunk size.

3. **Breakpoint-based semantic chunking:** Similar to the recursive approach, semantic chunking displayed a positive correlation between maximum chunk size and recall, with 0.52 being the maximum value reported by the configuration with $N = 1024$. For the similarity threshold percentile Q, a crossover effect was observed. For smaller maximum chunk sizes $N < 512$, the lower threshold $Q = 70$ yielded higher recall values. However, as the maximum chunk size increased, the higher threshold $Q = 90$ prevailed. The same relationship was also observed for the precision and IoU metrics. For precision$_\Omega$, $Q = 90$ yielded higher values at every N value.

4. **Hierarchical chunking:** At the smallest maximum chunk size $N = 128$, hierarchical chunking exhibited the highest precision (0.22), precision$_\Omega$ (0.17), and IoU (0.17) values of any tested configuration. In addition, it reported the second highest recall value at this N value with 0.32. However, increasing the chunk size did not lead to an increase in recall as it did for the other strategies, with recall values stagnating and even recessing as the maximum chunk length increased.

Across all reported configurations, we observe that the reported IoU values are highly correlated with the configurations precision. Especially for lower precision scores (Pr < 0.1), the IoU remained within 0.01 of the reported precision. The metric also responded more sensitively to changes in the precision value while staying stable as recall varied.

4.2.2. General Document QA Datasets

In addition to evaluating the chunking strategies on the manually annotated oncology QA pairs, we measured their performance on the document corpora provided by the research of B. Smith and Troynikov [58]. Specifically, five different corpora are provided, including chat dialogues, wikipedia excerpts, and Pubmed, which is comprised of biomedical journal literature [58]. We report both the results averaged across these corpora (Table 4.6) and the individual results on the Pubmed corpus (Table 4.7).

		N	IoU	Recall	Precision	Precision$_\Omega$
Fixed-Size	$O = 0$	128	$\underline{0.07 \pm 0.06}$	0.82 ± 0.33	0.07 ± 0.06	0.28 ± 0.14
	$O = 26$	128	$\underline{0.07 \pm 0.05}$	0.85 ± 0.31	$\underline{0.08 \pm 0.06}$	0.25 ± 0.12
	$O = 0$	256	0.04 ± 0.03	0.85 ± 0.32	0.04 ± 0.03	0.17 ± 0.09
	$O = 52$	256	0.04 ± 0.03	0.86 ± 0.32	0.04 ± 0.03	0.15 ± 0.09
	$O = 0$	512	0.02 ± 0.02	0.87 ± 0.31	0.02 ± 0.02	0.10 ± 0.06
	$O = 104$	512	0.02 ± 0.02	0.89 ± 0.30	0.02 ± 0.02	0.09 ± 0.06
	$O = 0$	1024	0.01 ± 0.01	0.87 ± 0.33	0.01 ± 0.01	0.06 ± 0.04
	$O = 208$	1024	0.01 ± 0.01	0.88 ± 0.32	0.01 ± 0.01	0.05 ± 0.04
Recursive	$O = 0$	128	$\mathbf{0.09 \pm 0.07}$	0.82 ± 0.35	$\mathbf{0.09 \pm 0.08}$	$\mathbf{0.40 \pm 0.21}$
	$O = 26$	128	$\mathbf{0.09 \pm 0.06}$	0.81 ± 0.35	$\mathbf{0.09 \pm 0.07}$	0.29 ± 0.14
	$O = 0$	256	0.05 ± 0.04	0.87 ± 0.32	0.05 ± 0.04	0.24 ± 0.17
	$O = 52$	256	0.05 ± 0.04	0.87 ± 0.31	0.05 ± 0.04	0.17 ± 0.10
	$O = 0$	512	0.03 ± 0.02	$\mathbf{0.91 \pm 0.29}$	0.03 ± 0.02	0.13 ± 0.09
	$O = 104$	512	0.02 ± 0.02	0.87 ± 0.32	0.02 ± 0.02	0.09 ± 0.06
	$O = 0$	1024	0.01 ± 0.01	$\underline{0.90 \pm 0.30}$	0.01 ± 0.01	0.06 ± 0.05
	$O = 208$	1024	0.01 ± 0.01	0.89 ± 0.31	0.01 ± 0.01	0.05 ± 0.04
Semantic	$Q = 70$	128	$\mathbf{0.09 \pm 0.07}$	0.82 ± 0.35	$\mathbf{0.09 \pm 0.07}$	$\underline{0.39 \pm 0.20}$
	$Q = 90$	128	$\mathbf{0.09 \pm 0.07}$	0.80 ± 0.36	$\mathbf{0.09 \pm 0.08}$	$\underline{0.39 \pm 0.21}$
	$Q = 70$	256	0.05 ± 0.04	0.87 ± 0.32	0.05 ± 0.04	0.24 ± 0.16
	$Q = 90$	256	0.05 ± 0.04	0.85 ± 0.34	0.05 ± 0.04	0.24 ± 0.17
	$Q = 70$	512	0.03 ± 0.02	0.89 ± 0.31	0.03 ± 0.02	0.13 ± 0.10
	$Q = 90$	512	0.03 ± 0.02	0.89 ± 0.31	0.03 ± 0.02	0.13 ± 0.10
	$Q = 70$	1024	0.01 ± 0.01	$\underline{0.90 \pm 0.30}$	0.01 ± 0.01	0.06 ± 0.05
	$Q = 90$	1024	0.01 ± 0.01	$\underline{0.90 \pm 0.30}$	0.01 ± 0.01	0.07 ± 0.05
Hierarchical	$B_h = 51$	128	$\mathbf{0.09 \pm 0.07}$	0.81 ± 0.35	$\mathbf{0.09 \pm 0.08}$	$\mathbf{0.40 \pm 0.21}$
	$B_h = 102$	256	0.05 ± 0.04	0.86 ± 0.33	0.05 ± 0.04	0.24 ± 0.17
	$B_h = 204$	512	0.03 ± 0.02	$\underline{0.90 \pm 0.30}$	0.03 ± 0.02	0.13 ± 0.09
	$B_h = 409$	1024	0.01 ± 0.01	$\underline{0.90 \pm 0.29}$	0.01 ± 0.01	0.06 ± 0.05

Table 4.6.: Results of the chunking evaluation averaged over all predefined corpora. All values are reported as $\mu \pm \sigma$. Highest μ are bolded and second-highest are underlined.

		N	IoU	Recall	Precision	Precision$_\Omega$
Fixed-Size	$O = 0$	128	0.08 ± 0.06	0.69 ± 0.39	0.08 ± 0.06	0.33 ± 0.15
	$O = 26$	128	0.08 ± 0.06	0.70 ± 0.37	0.08 ± 0.06	0.29 ± 0.12
	$O = 0$	256	0.05 ± 0.04	0.73 ± 0.39	0.05 ± 0.04	0.20 ± 0.09
	$O = 52$	256	0.05 ± 0.04	0.74 ± 0.40	0.05 ± 0.04	0.20 ± 0.11
	$O = 0$	512	0.03 ± 0.02	0.81 ± 0.36	0.03 ± 0.02	0.13 ± 0.07
	$O = 104$	512	0.03 ± 0.02	0.80 ± 0.38	0.03 ± 0.02	0.11 ± 0.07
	$O = 0$	1024	0.01 ± 0.01	0.82 ± 0.38	0.01 ± 0.01	0.07 ± 0.04
	$O = 208$	1024	0.01 ± 0.01	0.83 ± 0.38	0.01 ± 0.01	0.06 ± 0.04
Recursive	$O = 0$	128	$\mathbf{0.10 \pm 0.09}$	0.66 ± 0.43	$\mathbf{0.10 \pm 0.09}$	$\underline{0.47 \pm 0.22}$
	$O = 26$	128	$\underline{0.09 \pm 0.08}$	0.64 ± 0.41	$\underline{0.09 \pm 0.08}$	0.35 ± 0.14
	$O = 0$	256	0.06 ± 0.05	0.75 ± 0.40	0.06 ± 0.05	0.32 ± 0.20
	$O = 52$	256	0.06 ± 0.05	0.72 ± 0.39	0.06 ± 0.05	0.21 ± 0.11
	$O = 0$	512	0.04 ± 0.03	$\mathbf{0.85 \pm 0.35}$	0.04 ± 0.03	0.18 ± 0.13
	$O = 104$	512	0.03 ± 0.03	0.80 ± 0.38	0.03 ± 0.03	0.12 ± 0.07
	$O = 0$	1024	0.02 ± 0.01	0.82 ± 0.37	0.02 ± 0.01	0.09 ± 0.07
	$O = 208$	1024	0.02 ± 0.01	0.81 ± 0.38	0.02 ± 0.01	0.07 ± 0.04
Semantic	$Q = 70$	128	$\mathbf{0.10 \pm 0.09}$	0.64 ± 0.43	$\mathbf{0.10 \pm 0.09}$	$\underline{0.47 \pm 0.21}$
	$Q = 90$	128	$\mathbf{0.10 \pm 0.09}$	0.65 ± 0.43	$\mathbf{0.10 \pm 0.09}$	$\underline{0.47 \pm 0.22}$
	$Q = 70$	256	0.06 ± 0.05	0.77 ± 0.40	0.06 ± 0.05	0.31 ± 0.19
	$Q = 90$	256	0.06 ± 0.05	0.74 ± 0.40	0.06 ± 0.05	0.31 ± 0.20
	$Q = 70$	512	0.03 ± 0.03	0.83 ± 0.36	0.03 ± 0.03	0.17 ± 0.12
	$Q = 90$	512	0.04 ± 0.03	0.83 ± 0.37	0.04 ± 0.03	0.18 ± 0.14
	$Q = 70$	1024	0.02 ± 0.01	$\underline{0.84 \pm 0.35}$	0.02 ± 0.01	0.09 ± 0.07
	$Q = 90$	1024	0.02 ± 0.01	0.83 ± 0.36	0.02 ± 0.01	0.09 ± 0.06
Hierarchical	$B_h = 51$	128	$\mathbf{0.10 \pm 0.09}$	0.65 ± 0.43	$\mathbf{0.10 \pm 0.09}$	$\mathbf{0.48 \pm 0.22}$
	$B_h = 102$	256	0.06 ± 0.05	0.75 ± 0.40	0.06 ± 0.05	0.32 ± 0.21
	$B_h = 204$	512	0.04 ± 0.03	$\mathbf{0.85 \pm 0.35}$	0.04 ± 0.03	0.18 ± 0.13
	$B_h = 409$	1024	0.02 ± 0.01	0.83 ± 0.37	0.02 ± 0.01	0.09 ± 0.07

Table 4.7.: Results of the chunking evaluation on the predefined PubMed corpus. All values are reported as $\mu \pm \sigma$. Highest μ are bolded and second-highest are underlined.

5. Discussion

We reflect on and interpret the results of our experiments with a primary focus on the suitability of established datasets and metrics for the oncology domain. We connect our findings to the research questions defined in Section 1.2 and identify the most promising methods across the different document segmentation tasks.

5.1. Document Layout Analysis Evaluation

A critical finding in the DLA evaluation was the inability of Document AI and LlamaParse, the two commercial cloud-based DP services, to correctly identify "list" elements. These failures and the overall low F1 scores across this category can likely be attributed to the visual similarities between list items and paragraphs. The parsing output shows that models often misclassify list items as paragraphs, interpreting the bullet points as additional characters within the element's text. The variability in the styling of bullet point markers is likely further complicating this distinction. This misclassification obscures the relationships between the list items, leading to the loss of structural information. This is especially problematic for the parsing of oncology guidelines, as lists often contain causal relationships such as the different steps of a specific therapy process.

We found that, while Docling's and MinerU's pipeline-based approaches produced similar results, there was a significant gap between the performance of their VLM-based counterparts. MinerU 2.5 VLM demonstrated that domain-specific VLMs can outperform general-purpose cloud-based models, such as Gemini 2.5 Flash, on DLA tasks. In contrast, Granite Docling's subpar performance suggests that its compact 256-million parameter architecture lacks the representational capacity to generalize to the scientific layouts in the PubLayNet dataset. While lightweight models offer efficiency benefits, our results highlight the need for a minimum level of model complexity to perform robust DLA on scientific layouts.

Besides Granite Docling, LlamaParse reported the lowest weighted F1 score. One possible factor that contributed to this performance slide is that LlamaParse occasionally returned invalid bounding boxes, including coordinates that were negative or extended beyond the page dimensions. This underperformance highlights the inherent risk of "hallucinations" that comes with VLM-based DP approaches. Besides its aforementioned inability to detect "list" elements, we found that LlamaParse was not able to correctly identify elements of the "figure" type. While the low F1 scores in the "figure" category reported by multiple other implementations can largely be attributed to the merging of adjacent plots and subplots into a single element, LlamaParse's failure is of a different nature. Specifically, it stems from a mismatch in recognized element types, with LlamaParse critically missing a category for

figures and images, making the recognition of these elements impossible. LlamaParse's unpredictability and overall disappointing performance in our DLA evaluation contrast its marketing and raise questions regarding its overall efficacy.

The observed disparities between the reported F1@50 and F1@50:95 scores highlight the difference in precision between the bounding boxes of different implementations. Increasing the IoU value yields a stricter metric, requiring predictions to be more closely aligned with the GTBBs in order to be counted as a TP.

We observed that the most pronounced decrease in F1 scores occurred within the "title" category. This sensitivity is inherently linked to the characteristics of the title elements' bounding boxes. As titles typically possess the smallest spatial area, minor deviations in the DTBB are enough to result in a significant drop in IoU. On the other hand, larger elements, such as tables, benefit from an increased absolute pixel tolerance which results in F1 scores remaining stable as the IoU threshold increases.

Across the evaluated DP approaches, MinerU 2.5 VLM experienced the largest decline when transitioning to the F1@50:95. This suggests that while the model excels at predicting correct labels and finding bounding boxes for each element, the precision of these bounding boxes lags behind that of other implementations. The opposite can be said about Granite Docling, which showed the smallest performance decrease, albeit as the model with the lowest weighted F1 score. For a RAG-based knowledge assistant that provides visual source attribution, a high F1@50:95 is critical, as inaccurate bounding boxes might result in the wrong text being highlighted for the user. This reduces overall transparency and trust in the system. Therefore, although MinerU 2.5 VLM produced the highest weighted F1@50 score, Docling with its specialized Heron DLA model is preferable for our application.

RQ1: We analyze how the challenges introduced by oncology guidelines are reflected specifically in the PubLayNet dataset. We find that PubLayNet incorporates many of the identified attributes of oncology guidelines, as the entirety of the dataset consists of medical journal literature with a text-centric layout in both double and single column variations. Additionally, the dataset contains large, complex tables and figures which is consistent with the CPGs. However, PubLayNet only consists of vertical documents, missing the variations in page orientations we identified in Section 2.1. Furthermore, a major drawback of PubLayNet and other established DLA datasets is the sole use of single-page PDF documents, making evaluations of multi-page structural elements, such as the long tables found in many oncology guidelines, impossible. Lastly, PubLayNet consists of scanned PDF documents, while CPGs are usually born-digital documents. This could lead to DP systems performing slightly worse on PubLayNet as they are not able to extract programmatic information from the document.

RQ2: We conclude that the F1 score is a suitable metric for measuring the DLA performance of the evaluated DP systems. Due to its invariance towards confidence scores, it enables the comparison of hard predictors, which is a critical aspect in the evolving landscape of DP implementations. Many of the evaluated implementations fall into this category, such as Docling, which does not include the element-level confidence scores in its final output. Most

importantly this enables the evaluation of the DLA performance for end-to-end VLMs, where reported confidence scores are often unreliable [112]. Through the F1 score we are able to measure the DLA performance of the entire DP system, taking additional post-processing steps into account that are performed after the DLA stage. However, we note that reporting the F1 score at a single fixed IoU threshold is insufficient to facilitate an in-depth analysis of the DLA performance and might mask underlying problems of the implementation's predictions. By determining the F1 scores at multiple IoU thresholds, we are able to draw conclusions about both the overall performance of the model and the accuracy of its bounding boxes. Lastly, reporting the metric individually for each category aids with identifying weak points of the DP approach, specifically for less common element types like "figure" or "list".

5.2. Content Extraction Evaluation

Our results show that VLMs have the potential to outperform pipeline-based approaches specifically for content extraction. MinerU 2.5 VLM dominating the benchmark and Gemini 2.5 Flash reaching a higher score than Google's specialized DP service Document AI LayoutParser confirmed this finding. However, Granite Docling's underwhelming performance again marks the critical need for a minimum amount of parameters for VLMs to be applicable in DP applications. The strong performance of Unstructured.io and Docling on the structure-only TEDS shows the positive effect of including a specialized module for table structure recognition in pipeline-based approaches. In combination with their relatively weak performance on the TEDS metric and the normalized edit distance, we conclude that these approaches have outstanding capabilities for understanding the structure of the tables but are ultimately constrained by their lacking text extraction.

We also observe that open-source implementations are able to outperform proprietary solutions on content extraction, as the closed-source LlamaParse finished last, struggling with correctly identifying text content and reading order. Even Document AI, which we included to represent the gold standard of DP implementations, was not able to match the overall performance of multiple open-source approaches.

Through the assessment of the content extraction performance of the DP approaches, we are able to identify the multiple key findings for our research questions.

RQ1: We evaluated the alignment between the document attributes of OmniDocBench and the identified characteristics of the oncology guidelines. Although only approximately 20 percent of the English scientific literature subset is comprised of medical literature, with the remainder covering other domains such as engineering, chemistry, mathematics, and computer science, we argue that the achieved results remain highly applicable to the oncology guidelines domain. This is because the document pages adhere to the same text-centric single and double-column format as the CPGs. In addition, OmniDocBench extends this structural variety by including triple-column layouts and irregular pages with sparse texts or figures. This additional variety further examines the stability of the DP approaches' content extraction capabilities. The inclusion of a large amount of complex tables in both vertical and horizontal

orientations also aligns well with the elements found in oncology guidelines. However, we note that OmniDocBench suffers from the same inherent limitations as PubLayNet. Its reliance on single-page PDF documents prevents the evaluation of multi-page structural elements, while the absence of digital-born documents leads to less favorable conditions than the oncology guidelines. Finally, while we acknowledge the limited scale of the evaluation subset, we argue that the high density of complex layouts and elements within these cases provides a rigorous stress test that measures the ability of the approaches to adapt to challenging compositions.

RQ2: We find that evaluating the efficacy of the DP implementations' content extraction capabilities requires an evaluation across multiple dimensions. Relying on a single metric for this evaluation would risk masking critical failures in the extraction process, while only giving limited insights into the element-specific extraction capabilities of the implementation.

To evaluate the accuracy of the implementations' text extraction we adopt the normalized edit distance. In contrast to n-gram based metrics which are also frequently used for this evaluation, such as BLEU [113], ROUGE [114], and METEOR [115], the normalized edit distance measures the character-level accuracy of the extracted text while strictly penalizing incorrect word order. We argue that in the context of our application domain, where character-level precision and correct word ordering are crucial to maintaining the scientific integrity of the medical recommendations, the normalized edit distance is the superior metric for measuring the quality of the extracted content.

While the normalized edit distance quantifies the accuracy of text within individual elements, it fails to assess the overall structure of the document. A DP implementation might extract the content of every detected element perfectly, but fail to arrange them in the correct order by missing the double-column layout of the document. The reading order edit distance fills this gap in the evaluation methodology by measuring the system's ability to preserve the logical sequence of the structural elements.

Lastly, we integrate the TEDS metric into our evaluation, as providing accurate content extraction for complex table elements is a crucial requirement for our application. In addition, we are able to analyze the models table structure understanding in isolation from its text extraction capabilities by reporting the metric's structure-only version.

5.3. Chunking Evaluation

Our evaluation of the chunking configurations revealed a consistent, overarching trade-off between token-wise precision and recall. As the maximum chunk size increased, recall generally improved while precision continuously degraded. The underlying explanation for this effect is straightforward. Larger chunks incorporate more tokens and are naturally more likely to contain a relevant text excerpt. However, because the absolute number of relevant tokens remains constant, expanding the chunk size also introduces additional noise, decreasing the overall ratio of relevant excerpts.

We find that there is a definitive upper limit for the positive correlation between chunk size and recall. For fixed-size chunking we observed that increasing N above 512 led to a decrease in overall recall. We hypothesize that extracting excessively large chunks of text without regarding the structure of the document introduces substantial noise. Consequently, the retriever is not able to locate the relevant information within the chunks, causing a decrease in retrieval performance. Additionally, our results show that including an overlap is less effective for recursive character chunking, leading to a decrease in precision with very little change in the recall value. For the general QA dataset, we even find that including an overlap led to a decrease in recall for the recursive character chunking. We therefore conclude that when chunks follow the structure of the document, overlap mainly introduces additional noise and should therefore only be included for chunking strategies that produce window passages.

Another crucial finding of our evaluation is the stagnation of the recall scores across different hierarchical chunking configurations. Unlike the other evaluated strategies, increasing the maximum chunk size did not lead to a change in recall for hierarchical chunking. Hereby, we note that the perceived increased performance of the strategy on the general QA datasets is not indicative for the performance of the hierarchical chunker itself. These datasets do not have any inherent document structure, which leads to the hierarchical chunker having to fall back exclusively on the recursive character chunker. Therefore, the results achieved by the hierarchical chunker mirror those of the recursive character chunker.

We hypothesize that this observed stagnation effect is caused by the headings which are prepended to the chunk's content. If multiple chunks begin with the same text, their resulting vector embeddings might become overly similar, preventing the retriever from successfully identifying the most relevant excerpts. However, because this strategy initially showed potential for yielding chunks with a high token-wise precision, we believe that exploring techniques to resolve this conflict represents a valuable direction for future research.

We attribute the crossover effect identified in breakpoint-based semantic chunking to the interaction between the similarity threshold and the maximum chunk size. Higher thresholds require more significant topical shifts, resulting in larger, semantically distinct chunks. Conversely, smaller thresholds produce more granular chunks. As the maximum chunk size increases, chunks from lower thresholds need to be merged to satisfy the minimum chunk length requirement. This merging process likely decreases semantic coherence inside the merged chunk. Therefore, we observe that configurations with higher thresholds tend to perform better as maximum chunk size increases. Regardless of the chosen parameters, our findings align with those of R. Qu, Tu, and Bao [12], finding that breakpoint-based semantic chunking does not justify its substantial computational overhead, as it does not yield any notable performance improvements compared to recursive character chunking.

Ultimately, while the evaluated chunking strategies demonstrated more comparable performance levels than initially expected, our findings identify recursive character chunking as the most robust solution for our application. Specifically, the configuration with a maximum chunk size of 512 tokens and zero overlap achieved high recall alongside promising precision values. Our overarching conclusion regarding the results of our evaluation is that tuning the

parameters can be more crucial than the selection of the strategy itself.

Based on the results of our evaluation, we were able to identify multiple key findings regarding which metrics served useful for evaluating chunking configurations.

RQ2: Completing our evaluation of suitable metrics, we identify the token-wise recall as the most important metric for measuring the efficacy of the chunking strategies. If the retrieval concludes with low recall, the generator will not have access to critical information needed to answer the query. This leads to inaccurate and ill-advised answers from the system that could potentially lead to dubious medical recommendations. Simultaneously, token-wise precision remains a critical aspect of our analysis. Low precision indicates an increased ratio of distracting information included in the chunks, which can subsequently deteriorate the performance of the generator [116, 105]. Additionally, low precision is detrimental to the accuracy of visual source attribution. If only a small amount of the highlighted text is relevant to the query, additional cognitive strain is placed on the user who has to manually review the content of the bloated bounding boxes. We find that the evaluation of different chunking strategies needs to take both metrics into account, albeit with a focus on maintaining a high token-wise recall.

We note that the inherent difficulty of the queries and the varying amount of required context to answer them currently suppress the achieved evaluation scores. We hypothesize that, with further retrieval optimization, both precision and recall could be drastically improved. Therefore, we find that $precision_\Omega$ proves to be an important indicator for the potential precision of the chunking configuration by establishing an upper bound for the retriever's precision. This metric essentially measures how accurately the created chunks align with the semantical structure of the document, while controlling for the performance of the retriever.

On the other hand, our results indicate that the token-wise IoU is not a suitable metric for this evaluation, as it is heavily biased towards the achieved precision score. Fundamentally, the IoU does not differentiate between the two bounding boxes when determining their union. For the token-wise IoU, this means that retrieving an irrelevant token is penalized the same as not retrieving a relevant token. We highlight the issues of this metric through a visual example in Figure 5.1. As recall is the most critical aspect of our evaluation, valuing it the same as precision does not reflect the nature of the retrieval process. While the generator is able to tolerate some noise [116], missing important information that is needed to answer the query can not be made up for in subsequent stages.

5.4. Document Segmentation Enhancements

Addressing **RQ3**, which investigates how document segmentation methods can be adapted and expanded on to fulfill the requirements of a RAG-based knowledge assistant, we integrated two distinct enhancements into our document segmentation pipeline. Firstly, we enhanced the relation integration of the DP implementations by implementing additional post-processing steps. Specifically, we filter out extraneous elements, infer the document's
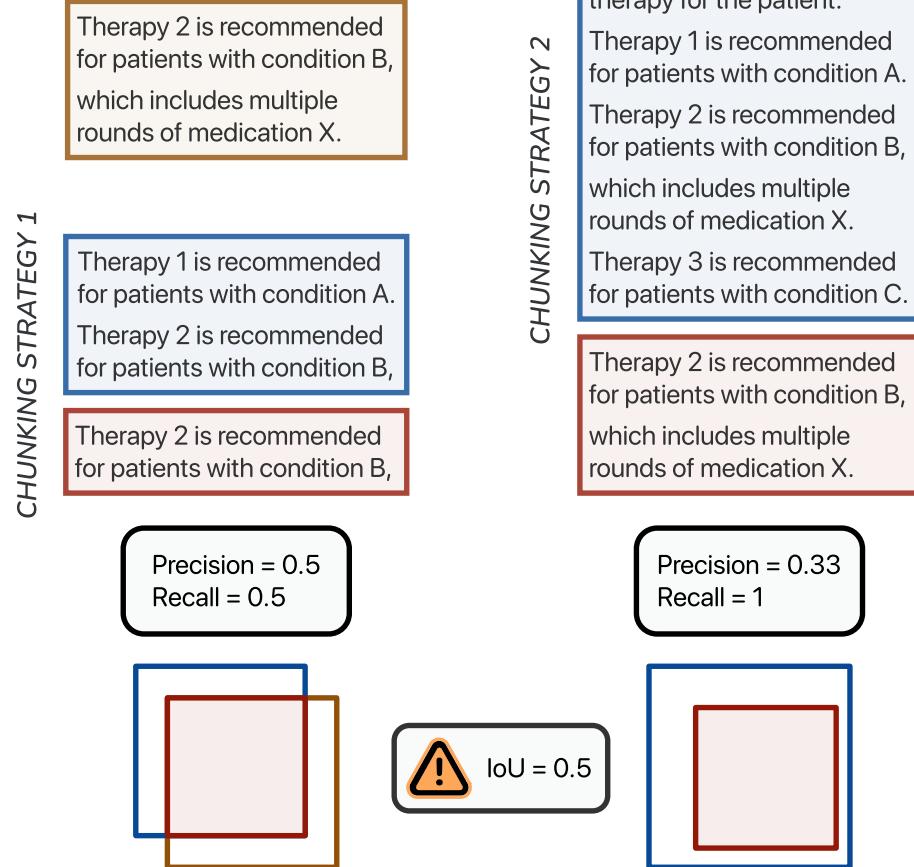
Figure 5.1.: Visual example for the token-wise IoU of two different chunking strategies. For a given query and chunking strategy, the tokens from the blue area are retrieved, while the tokens from the yellow area are needed to answer the question. The set of correctly retrieved tokens is denoted by the red area. We see that despite the chunk from chunking strategy 2 providing the entirety of the needed information needed to answer a question, the IoU score is the same as strategy 1 due to the lower precision of strategy 2.

structural hierarchy based on the reading order, and determine span-level bounding boxes. These additions optimize the document tree, enabling the use of structure-dependent chunking strategies, such as hierarchical chunking, while ensuring that downstream chunking output only contains relevant content. Our second contribution entails the proposition of a novel token-based methodology for the chunking process. This approach enables traceability for every constituent token, even when employing chunking strategies that do not inherently respect the structure of the document. Combined with the span-level bounding boxes introduced in the DP post-processing, our pipeline successfully generates chunk bounding boxes at a higher level of granularity than previous implementations. However, it is important to note that the precision of the produced bounding boxes has so far only been verified through qualitative, small-scale evaluations. Therefore, future work should incorporate a more rigorous quantitative evaluation to validate the robustness of this novel approach.

5.5. Key Findings and Best Practices

We summarize the findings described in this chapter for each of the research questions, giving a list of actionable recommendations for future applications of the research conducted in this study.

RQ1: How are the challenges introduced by oncology guidelines reflected in established benchmarks for document parsing?

We found that established benchmarks and datasets generally focus on a single aspect of the DP process. PubLayNet proved to be valuable for evaluating the implementations' DLA capabilities on documents that are from a similar medical domain as the guideline documents. OmniDocBench provides an extensive stress test that evaluates the content extraction performance on a range of challenging scientific layouts, albeit with a very small sample size. However, we found that some attributes of CPGs are not reflected, namely multi-page structural elements and born-digital PDF files. We conclude that there is still a need for future research regarding the creation of more complete datasets.

RQ2: Which metrics are most useful for measuring the effectiveness of document parsing and chunking methods?

We identified a set of valuable metrics that can be applied during different stages of the document segmentation process.

For DLA, the (weighted) F1 score is useful for comparing the performance of the different implementations, regardless of whether they include trustworthy confidence scores in their output. We recommend evaluating the scores at different IoU thresholds to gain insights into the overall quality of the produced bounding boxes. We identified F1@50:95 as the most useful metric to rank a group of DP implementations, as it rewards high-precision localizations while still acknowledging acceptable predictions.

For content extraction, we recommend using a range of metrics to measure the performance of different aspects of the process. Specifically, the normalized edit distance is useful for measuring the text extraction performance and can also be applied for evaluating the correctness of the reading order. For table structure recognition, we recommend TEDS and especially its structure-only variant to measure the quality of the implementation's table structure.

We argue that token-wise recall is the most important metric for measuring the quality of the chunking process. The produced chunks should also have at least a minimum level of token-wise precision to ensure that the generator will be able to find the relevant context inside the retrieved information. For experiments with little retrieval optimization, $precision_\Omega$ proved useful to identify how similar the produced chunks are to the ground truth. We find that combining recall and precision into a single metric is not recommended due to the imbalanced relevancy of the two metrics.

> **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?

We find that pipeline-based approaches, specifically MinerU 2.5 Pipeline and Docling, performed best for DLA, while both MinerU 2.5 VLM and Pipeline are the superior choices for content extraction. MinerU 2.5 VLM's slow parsing speed likely makes it infeasible for the project, leaving MinerU 2.5 Pipeline and Docling as the optimal choices. Docling provides more granular table elements, processes documents significantly faster, and has a very permissive licensing. MinerU provides better content extraction and span-level bounding boxes for text elements.

The evaluation of the chunking strategies revealed that recursive character chunking performed best among the tested configurations. For RAG applications building on oncology guideline documents, we recommend starting with an initial chunk size of 512 tokens and performing further parameter tuning.

Our improvements to the DP process include the filtering of extraneous elements and optimization of the document tree. For chunking, we introduce a novel methodology that links chunk tokens to the structural element they belong to. This enables visual source attribution with high-granularity bounding boxes.

6. Conclusion

This study encompasses an in-depth comparative analysis of current document segmentation configurations and their application for oncology guideline documents. We developed a modular document segmentation pipeline, proposing a unified methodology that addresses the fragmented landscape of available DP implementations. Furthermore, motivated by the lack of traceability of current solutions, we introduced a novel approach to the chunking problem that prioritizes traceability and enables visual source attribution for established strategies. In total, we provide integrations for eight different DP implementations and four chunking strategies, facilitating a rigorous evaluation across a multitude of possible module combinations.

By critically examining existing benchmarks, we established a set of evaluation metrics and datasets tailored to the requirements of the oncology guideline documents. After reviewing the capabilities of the DP approaches and 28 distinct chunking configurations, our findings indicate that, while VLM-based approaches demonstrate significant potential for DP, especially for content extraction, they remain limited by their output variability and positional inaccuracy. Consequently, we conclude that current pipeline-based approaches, specifically MinerU 2.5 Pipeline and Docling, offer the stability and completeness required for the data preparation of the RAG-based medical knowledge assistant. Our evaluation of the chunking techniques showed that there is no indication that complex breakpoint-based semantic chunking warrants its increased computational requirements. Instead, we come to the conclusion that recursive character chunking, if configured correctly, is likely to provide the best performance for the oncology guideline domain.

As the field of DP is continuously evolving, a primary focus in the development of the document segmentation pipeline was to ensure a modular architecture that supports the integration of emerging techniques. We therefore encourage future research towards exploring novel DP approaches and applying our proposed methodology to additional chunking techniques.

Another direction that could be explored in future experiments involves addressing the limitations of our current evaluation framework. In particular, expanding the performance assessment of DP implementations on multi-page PDF documents would bridge a significant gap in the existing research landscape.

A. General Addenda

Listing A.1: Prompt to apply the Gemini 2.5 Flash model to the DP task. Gemini models are trained to output coordinates from 0 to 1000, with the origin at the top-left 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 the 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

List of Tables

Acronyms

AI artificial intelligence. 1, 11, 20, 22

AP average precision. 26, 27

API application programming interface. 21

AWMF Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften. 4, 40

CNN convolutional neural network. 10

CPG clinical practice guidelines. 1, 4, 5, 9, 29, 46, 47, 52

DLA document layout analysis. v, 10, 11, 16, 19–21, 26, 29, 34, 36, 39, 45–47, 52, 53

DP document parsing. v, 2, 3, 9–11, 13–22, 26, 28–32, 35–37, 39, 40, 45–48, 50, 52–55, 57, 58

DTBB detected bounding box. 27, 46

ESMO European Society for Medical Oncology. 4

FN false negative. 27

FP false positive. 27

GLD Generalized Levenshtein Distance. 30–32

GOT General OCR Theory. 11

GTBB ground truth bounding box. 27, 28, 46

HTML Hypertext Markup Language. 11, 29, 31

IoU intersection over union. v, 27, 28, 34, 36, 42, 46, 47, 50–52, 59

JSON JavaScript Object Notation. 13, 14, 19, 55

LLM large language model. 1, 5, 7–9, 13, 22, 32

LTRB left-top-right-bottom. 7, 16

NCCN National Comprehensive Cancer Network. 1, 4

NLP natural language processing. 1, 5, 6

NLTK Natural Language Toolkit. 24

OCR optical character recognition. 10, 11, 18–20, 22

PDF portable document format. 1, 4, 9, 10, 14, 16–20, 29, 38, 39, 46, 48, 52, 54

PMCOA PubMed Central Open Access. 26

QA question-answer. 32, 33, 40–42, 49, 60

RAG retrieval-augmented generation. v, 1, 2, 7–9, 12, 15, 17, 19, 21, 22, 24, 32, 35, 46, 50, 53, 54, 59

TEDS Tree-Edit-Distance-based Similarity. v, 31, 38, 47, 48, 53

TP true positive. 27, 28, 34, 46

TUM Technical University of Munich. 32, 40

VLM vision-language model. 7, 10–12, 20, 21, 35, 36, 38, 39, 45–47, 53, 54

XML extensible markup language. 11, 13, 20

Bibliography

[1] E. Steinberg, S. Greenfield, D. M. Wolman, M. Mancher, and R. Graham. *Clinical practice guidelines we can trust*. national academies press, 2011.

[2] National Comprehensive Cancer Network. *About Clinical Practice Guidelines*. NCCN. URL: https://www.nccn.org/guidelines/guidelines-process/about-nccn-clinical-practice-guidelines (visited on 01/29/2026).

[3] B. H. Kann, S. B. Johnson, H. J. Aerts, R. H. Mak, and P. L. Nguyen. "Changes in length and complexity of clinical practice guidelines in oncology, 1996-2019". In: *JAMA Network Open* 3.3 (2020), e200841–e200841.

[4] Chair of Software Engineering for Business Information Systems, TUM School of Computation, Information and Technology, Technichal University of Munich. *AI-Based Knowledge Assistant for Cancer Care (Aidvice)*. 2025. URL: https://www.cs.cit.tum.de/sebis/research/natural-language-processing/ai-based-knowledge-assistant-for-cancer-care-aidvice/ (visited on 01/28/2026).

[5] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela. *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. 2021. arXiv: 2005.11401 [cs.CL]. URL: https://arxiv.org/abs/2005.11401.

[6] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, and H. Wang. *Retrieval-Augmented Generation for Large Language Models: A Survey*. 2024. arXiv: 2312.10997 [cs.CL]. URL: https://arxiv.org/abs/2312.10997.

[7] S. Wang, F. Zhao, D. Bu, and et al. "LINS: A general medical Q&A framework for enhancing the quality and credibility of LLM-generated responses". In: *Nature Communications* 16.1 (Oct. 2025), p. 9076. DOI: 10.1038/s41467-025-64142-2. URL: https://doi.org/10.1038/s41467-025-64142-2.

[8] J. Hladěna, K. Šteflovič, P. Čech, K. Štekerová, and A. Žváčková. "The Effect of Chunk Size on the RAG Performance". In: *Software Engineering: Emerging Trends and Practices in System Development*. Ed. by R. Silhavy and P. Silhavy. Cham: Springer Nature Switzerland, 2025, pp. 317–326. ISBN: 978-3-032-00712-4.

[9] T. Chen, H. Wang, S. Chen, W. Yu, K. Ma, X. Zhao, H. Zhang, and D. Yu. *Dense X Retrieval: What Retrieval Granularity Should We Use?* 2024. arXiv: 2312.06648 [cs.CL]. URL: https://arxiv.org/abs/2312.06648.

[10] L. Müller, J. Holstein, S. Bause, G. Satzger, and N. Kühl. *Data Quality Challenges in Retrieval-Augmented Generation*. 2025. arXiv: 2510.00552 [cs.AI]. URL: https://arxiv.org/abs/2510.00552.

[11] X. Ma, S. Zhuang, B. Koopman, G. Zuccon, W. Chen, and J. Lin. *VISA: Retrieval Augmented Generation with Visual Source Attribution*. 2024. arXiv: 2412.14457 [cs.IR]. URL: https://arxiv.org/abs/2412.14457.

[12] R. Qu, R. Tu, and F. Bao. "Is semantic chunking worth the computational cost?" In: *Findings of the Association for Computational Linguistics: NAACL 2025*. 2025, pp. 2155–2177.

[13] Q. Zhang, B. Wang, V. S.-J. Huang, J. Zhang, Z. Wang, H. Liang, C. He, and W. Zhang. *Document Parsing Unveiled: Techniques, Challenges, and Prospects for Structured Information Extraction*. 2025. arXiv: 2410.21169 [cs.MM]. URL: https://arxiv.org/abs/2410.21169.

[14] H. Chase. *LangChain*. https://github.com/langchain-ai/langchain. Oct. 2022. URL: https://github.com/langchain-ai/langchain.

[15] J. Liu. *LlamaIndex*. Nov. 2022. DOI: 10.5281/zenodo.1234. URL: https://github.com/jerryjliu/llama_index.

[16] N. Livathinos, C. Auer, M. Lysak, A. Nassar, M. Dolfi, P. Vagenas, C. B. Ramis, M. Omenetti, K. Dinkla, Y. Kim, S. Gupta, R. T. de Lima, V. Weber, L. Morin, I. Meijer, V. Kuropiatnyk, and P. W. J. Staar. *Docling: An Efficient Open-Source Toolkit for AI-driven Document Conversion*. 2025. arXiv: 2501.17887 [cs.CL]. URL: https://arxiv.org/abs/2501.17887.

[17] LangChain Inc. *LangChain Docs: Splitting recursively*. URL: https://docs.langchain.com/oss/python/integrations/splitters/recursive_text_splitter (visited on 01/27/2026).

[18] E. Guerra-Farfan, Y. Garcia-Sanchez, M. Jornet-Gibert, J. H. Nuñez, M. Balaguer-Castro, and K. Madden. "Clinical practice guidelines: The good, the bad, and the ugly". In: *Injury* 54 (2023). AOTrauma Europe Supplement: Clinical Research: Lessons Learned-Looking Ahead, S26–S29. ISSN: 0020-1383. DOI: https://doi.org/10.1016/j.injury.2022.01.047. URL: https://www.sciencedirect.com/science/article/pii/S0020138322000778.

[19] N. L. Stout, D. Santa Mina, K. D. Lyons, K. Robb, and J. K. Silver. "A systematic review of rehabilitation and exercise recommendations in oncology guidelines". In: *CA: a cancer journal for clinicians* 71.2 (2021), pp. 149–175.

[20] European Society for Medical Oncology. *European Society for Medical Oncology (ESMO)*. URL: https://www.esmo.org/ (visited on 02/04/2026).

[21] Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften. *Arbeitsgemeinschaft der Wissenschaftlichen Medizinischen Fachgesellschaften (AWMF)*. URL: https://www.awmf.org (visited on 02/04/2026).

[22] S. Banerjee, C. M. Booth, E. Bruera, M. W. Buechler, A. Drilon, T. J. Fry, I. M. Ghobrial, L. Gianni, R. K. Jain, G. Kroemer, et al. "Two decades of advances in clinical oncology—lessons learned and future directions". In: *Nature Reviews Clinical Oncology* 21.11 (2024), pp. 771–780.

[23] International Organization for Standardization. *ISO 32000-2:2020: Document management – Portable document format – Part 2: PDF 2.0*. Internation Organization for Standardization, 2020.

[24] J. Hirschberg and C. D. Manning. "Advances in natural language processing". In: *Science* 349.6245 (2015), pp. 261–266.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[26] B. Min, H. Ross, E. Sulem, A. P. B. Veyseh, T. H. Nguyen, O. Sainz, E. Agirre, I. Heintz, and D. Roth. "Recent Advances in Natural Language Processing via Large Pre-trained Language Models: A Survey". In: *ACM Comput. Surv.* 56.2 (Sept. 2023). ISSN: 0360-0300. DOI: 10.1145/3605943. URL: https://doi.org/10.1145/3605943.

[27] X. Song, A. Salcianu, Y. Song, D. Dopson, and D. Zhou. "Fast wordpiece tokenization". In: *Proceedings of the 2021 conference on empirical methods in natural language processing*. 2021, pp. 2089–2103.

[28] S. J. Mielke, Z. Alyafeai, E. Salesky, C. Raffel, M. Dey, M. Gallé, A. Raja, C. Si, W. Y. Lee, B. Sagot, and S. Tan. *Between words and characters: A Brief History of Open-Vocabulary Modeling and Tokenization in NLP*. 2021. arXiv: 2112.10508 [cs.CL]. URL: https://arxiv.org/abs/2112.10508.

[29] M. Schuster and K. Nakajima. "Japanese and Korean voice search". In: *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2012, pp. 5149–5152. DOI: 10.1109/ICASSP.2012.6289079.

[30] T. Kudo and J. Richardson. "SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Ed. by E. Blanco and W. Lu. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 66–71. DOI: 10.18653/v1/D18-2012. URL: https://aclanthology.org/D18-2012/.

[31] T. Kudo. "Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by I. Gurevych and Y. Miyao. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 66–75. DOI: 10.18653/v1/P18-1007. URL: https://aclanthology.org/P18-1007/.

[32] D. Jurafsky and J. H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*. 3rd. Online manuscript released January 6, 2026. 2026. URL: https://web.stanford.edu/~jurafsky/slp3/.

[33] A. Ghosh, A. Acharya, S. Saha, V. Jain, and A. Chadha. *Exploring the Frontier of Vision-Language Models: A Survey of Current Methodologies and Future Directions*. 2025. arXiv: 2404.07214 [cs.CV]. URL: https://arxiv.org/abs/2404.07214.

[34] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, et al. "Qwen2. 5-vl technical report". In: *arXiv preprint arXiv:2502.13923* (2025).

[35] L. d. F. D. Costa and R. M. Cesar Jr. *Shape analysis and classification: theory and practice.* CRC Press, Inc., 2000.

[36] A. Zhang, Z. C. Lipton, M. Li, and A. J. Smola. *Dive into Deep Learning*. 2023. arXiv: `2106.11342 [cs.LG]`. URL: `https://arxiv.org/abs/2106.11342`.

[37] N. Aksoy, Z. A. Güven, and M. O. Ünalır. "Understanding the Impact of Dataset Characteristics on RAG based Multi-hop QA Performance". In: (2025).

[38] J. Ni, C. Qu, J. Lu, Z. Dai, G. H. Abrego, J. Ma, V. Zhao, Y. Luan, K. Hall, M.-W. Chang, et al. "Large dual encoders are generalizable retrievers". In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 2022, pp. 9844–9855.

[39] T. Gao, H. Yen, J. Yu, and D. Chen. "Enabling Large Language Models to Generate Text with Citations". In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by H. Bouamor, J. Pino, and K. Bali. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 6465–6488. DOI: `10.18653/v1/2023.emnlp-main.398`. URL: `https://aclanthology.org/2023.emnlp-main.398/`.

[40] L. Ouyang, Y. Qu, H. Zhou, J. Zhu, R. Zhang, Q. Lin, B. Wang, Z. Zhao, M. Jiang, X. Zhao, J. Shi, F. Wu, P. Chu, M. Liu, Z. Li, C. Xu, B. Zhang, B. Shi, Z. Tu, and C. He. *OmniDocBench: Benchmarking Diverse PDF Document Parsing with Comprehensive Annotations*. 2024. arXiv: `2412.07626 [cs.CV]`. URL: `https://arxiv.org/abs/2412.07626`.

[41] B. Wang, C. Xu, X. Zhao, L. Ouyang, F. Wu, Z. Zhao, R. Xu, K. Liu, Y. Qu, F. Shang, B. Zhang, L. Wei, Z. Sui, W. Li, B. Shi, Y. Qiao, D. Lin, and C. He. *MinerU: An Open-Source Solution for Precise Document Content Extraction*. 2024. arXiv: `2409.18839 [cs.CV]`. URL: `https://arxiv.org/abs/2409.18839`.

[42] D. S. Team. *Docling Technical Report*. Tech. rep. Version 1.0.0. Aug. 2024. DOI: `10.48550/arXiv.2408.09869`. eprint: `2408.09869`. URL: `https://arxiv.org/abs/2408.09869`.

[43] H. Xing, F. Gao, Q. Zheng, Z. Zhu, Z. Shao, and M. Yan. "Intelligent Document Parsing: Towards End-to-end Document Parsing via Decoupled Content Parsing and Layout Grounding". In: *Findings of the Association for Computational Linguistics: EMNLP 2025*. Ed. by C. Christodoulopoulos, T. Chakraborty, C. Rose, and V. Peng. Suzhou, China: Association for Computational Linguistics, Nov. 2025, pp. 19987–19998. ISBN: 979-8-89176-335-7. DOI: `10.18653/v1/2025.findings-emnlp.1088`. URL: `https://aclanthology.org/2025.findings-emnlp.1088/`.

[44] J. Niu, Z. Liu, Z. Gu, B. Wang, L. Ouyang, Z. Zhao, T. Chu, T. He, F. Wu, Q. Zhang, Z. Jin, G. Liang, R. Zhang, W. Zhang, Y. Qu, Z. Ren, Y. Sun, Y. Zheng, D. Ma, Z. Tang, B. Niu, Z. Miao, H. Dong, S. Qian, J. Zhang, J. Chen, F. Wang, X. Zhao, L. Wei, W. Li, S. Wang, R. Xu, Y. Cao, L. Chen, Q. Wu, H. Gu, L. Lu, K. Wang, D. Lin, G. Shen, X. Zhou, L. Zhang, Y. Zang, X. Dong, J. Wang, B. Zhang, L. Bai, P. Chu, W. Li, J. Wu, L. Wu, Z. Li, G. Wang, Z. Tu, C. Xu, K. Chen, Y. Qiao, B. Zhou, D. Lin, W. Zhang, and C. He.

MinerU2.5: A Decoupled Vision-Language Model for Efficient High-Resolution Document Parsing. 2025. arXiv: `2509.22186` [`cs.CV`]. URL: `https://arxiv.org/abs/2509.22186`.

[45] Unstructured.io Team. *Unstructured.io: Open-Source Pre-Processing Tools for Unstructured Data*. URL: `https://unstructured.io` (visited on 01/20/2026).

[46] Z. Li, Y. Liu, Q. Liu, Z. Ma, Z. Zhang, S. Zhang, Z. Guo, J. Zhang, X. Wang, and X. Bai. *MonkeyOCR: Document Parsing with a Structure-Recognition-Relation Triplet Paradigm*. 2025. arXiv: `2506.05218` [`cs.CV`]. URL: `https://arxiv.org/abs/2506.05218`.

[47] N. Livathinos, C. Auer, A. Nassar, R. T. de Lima, M. Lysak, B. Ebouky, C. Berrospi, M. Dolfi, P. Vagenas, M. Omenetti, K. Dinkla, Y. Kim, V. Weber, L. Morin, I. Meijer, V. Kuropiatnyk, T. Strohmeyer, A. S. Gurbuz, and P. W. J. Staar. *Advanced Layout Analysis Models for Docling*. 2025. arXiv: `2509.11720` [`cs.CV`]. URL: `https://arxiv.org/abs/2509.11720`.

[48] T. Sun, C. Cui, Y. Du, and Y. Liu. *PP-DocLayout: A Unified Document Layout Detection Model to Accelerate Large-Scale Data Construction*. 2025. arXiv: `2503.17213` [`cs.CV`]. URL: `https://arxiv.org/abs/2503.17213`.

[49] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi. "You Only Look Once: Unified, Real-Time Object Detection". In: *CoRR* abs/1506.02640 (2015). arXiv: `1506.02640`. URL: `http://arxiv.org/abs/1506.02640`.

[50] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. "End-to-End Object Detection with Transformers". In: *CoRR* abs/2005.12872 (2020). arXiv: `2005.12872`. URL: `https://arxiv.org/abs/2005.12872`.

[51] Y. Huang, T. Lv, L. Cui, Y. Lu, and F. Wei. *LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking*. 2022. arXiv: `2204.08387` [`cs.CL`]. URL: `https://arxiv.org/abs/2204.08387`.

[52] N. Islam, Z. Islam, and N. Noor. "A survey on optical character recognition system". In: *arXiv preprint arXiv:1710.05703* (2017).

[53] R. Kittinaradorn and JaidedAI. *EasyOCR*. `https://github.com/JaidedAI/EasyOCR`. 2020.

[54] R. Smith. "An overview of the Tesseract OCR engine". In: *Ninth international conference on document analysis and recognition (ICDAR 2007)*. Vol. 2. IEEE. 2007, pp. 629–633.

[55] Unstructured.io Team. *Unstructured.io: Documentation for the open-source library*. URL: `https://docs.unstructured.io/open-source/introduction/overview` (visited on 01/28/2026).

[56] H. Wei, C. Liu, J. Chen, J. Wang, L. Kong, Y. Xu, Z. Ge, L. Zhao, J. Sun, Y. Peng, C. Han, and X. Zhang. *General OCR Theory: Towards OCR-2.0 via a Unified End-to-end Model*. 2024. arXiv: `2409.01704` [`cs.CV`]. URL: `https://arxiv.org/abs/2409.01704`.

[57] Y. Li, G. Yang, H. Liu, B. Wang, and C. Zhang. *dots.ocr: Multilingual Document Layout Parsing in a Single Vision-Language Model*. 2025. arXiv: `2512.02498` [`cs.CV`]. URL: `https://arxiv.org/abs/2512.02498`.

[58] B. Smith and A. Troynikov. *Evaluating Chunking Strategies for Retrieval*. Tech. rep. Chroma, June 2024. URL: https://research.trychroma.com/evaluating-chunking.

[59] K. Kise, M. Junker, A. Dengel, and K. Matsumoto. "Passage-Based Document Retrieval as a Tool for Text Mining with User's Information Needs". In: *Discovery Science*. Ed. by K. P. Jantke and A. Shinohara. Berlin, Heidelberg: Springer Berlin Heidelberg, 2001, pp. 155–169. ISBN: 978-3-540-45650-6.

[60] J. P. Callan. "Passage-level evidence in document retrieval". In: *SIGIR'94: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, organised by Dublin City University*. Springer. 1994, pp. 302–310.

[61] LlamaIndex. *NodeParser Modules*. URL: https://developers.llamaindex.ai/python/framework/module_guides/loading/node_parsers/modules/ (visited on 01/27/2026).

[62] S. Jaiswal, P. Bisht, K. Kansara, and M. S. Datta. "Comparison of Chunking Techniques Across Diverse Document Types in NLP Retrieval Tasks". In: *2025 International Conference on Responsible, Generative and Explainable AI (ResGenXAI)*. 2025, pp. 1–6. DOI: 10.1109/ResgenXAI64788.2025.11344045.

[63] O. Koshorek, A. Cohen, N. Mor, M. Rotman, and J. Berant. "Text segmentation as a supervised learning task". In: *arXiv preprint arXiv:1803.09337* (2018).

[64] A. V. Duarte, J. D. Marques, M. Graça, M. Freire, L. Li, and A. L. Oliveira. "Lumber-chunker: Long-form narrative document segmentation". In: *Findings of the Association for Computational Linguistics: EMNLP 2024*. 2024, pp. 6473–6486.

[65] C. He, W. Li, Z. Jin, C. Xu, B. Wang, and D. Lin. "Opendatalab: Empowering general artificial intelligence with open datasets". In: *arXiv preprint arXiv:2407.13773* (2024).

[66] Google Cloud. *Document AI: Process documents with Gemini layout parser*. URL: https://docs.cloud.google.com/document-ai/docs/layout-parse-chunk (visited on 02/06/2026).

[67] J. X. M. Artifex Software Inc. *PyMuPDF*. Version 1.26.7. 2025. URL: https://github.com/pymupdf/PyMuPDF.

[68] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun. *YOLOX: Exceeding YOLO Series in 2021*. 2021. arXiv: 2107.08430 [cs.CV]. URL: https://arxiv.org/abs/2107.08430.

[69] A. Nassar, N. Livathinos, M. Lysak, and P. Staar. "TableFormer: Table Structure Understanding With Transformers". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2022, pp. 4614–4623. DOI: https://doi.org/10.1109/CVPR52688.2022.00457.

[70] Y. Zhao, W. Lv, S. Xu, J. Wei, G. Wang, Q. Dang, Y. Liu, and J. Chen. *DETRs Beat YOLOs on Real-time Object Detection*. 2024. arXiv: 2304.08069 [cs.CV]. URL: https://arxiv.org/abs/2304.08069.

[71] B. Pfitzmann, C. Auer, M. Dolfi, A. S. Nassar, and P. W. J. Staar. "DocLayNet: A Large Human-Annotated Dataset for Document-Layout Analysis". In: (2022). DOI: 10.1145/3534678.353904. URL: https://arxiv.org/abs/2206.01062.

[72] IBM Research. *Granite Docling Documentation*. IBM. URL: https://www.ibm.com/granite/docs/models/docling (visited on 02/03/2026).

[73] A. Nassar, A. Marafioti, M. Omenetti, M. Lysak, N. Livathinos, C. Auer, L. Morin, R. T. de Lima, Y. Kim, A. S. Gurbuz, M. Dolfi, M. Farré, and P. W. J. Staar. *SmolDocling: An ultra-compact vision-language model for end-to-end multi-modal document conversion*. 2025. arXiv: 2503.11576 [cs.CV]. URL: https://arxiv.org/abs/2503.11576.

[74] Z. Zhao, H. Kang, B. Wang, and C. He. *DocLayout-YOLO: Enhancing Document Layout Analysis through Diverse Synthetic Data and Global-to-Local Adaptive Perception*. 2024. arXiv: 2410.12628 [cs.CV]. URL: https://arxiv.org/abs/2410.12628.

[75] B. Wang, Z. Gu, G. Liang, C. Xu, B. Zhang, B. Shi, and C. He. *UniMERNet: A Universal Network for Real-World Mathematical Expression Recognition*. 2024. arXiv: 2404.15254 [cs.CV]. URL: https://arxiv.org/abs/2404.15254.

[76] G. Comanici, E. Bieber, M. Schaekermann, et al. *Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities*. 2025. arXiv: 2507.06261 [cs.CL]. URL: https://arxiv.org/abs/2507.06261.

[77] Google. *Image Understanding*. Google AI for Developers. URL: https://ai.google.dev/gemini-api/docs/image-understanding (visited on 02/03/2026).

[78] LlamaIndex. *LlamaParse: GenAI-native document parsing platform*. 2024. URL: https://www.llamaindex.ai/llamaparse (visited on 02/03/2026).

[79] Google Cloud. *Document AI: Enterprise Document OCR*. URL: https://docs.cloud.google.com/document-ai/docs/enterprise-document-ocr (visited on 02/06/2026).

[80] N. Reimers and I. Gurevych. "Sentence-bert: Sentence embeddings using siamese bert-networks". In: *arXiv preprint arXiv:1908.10084* (2019).

[81] T. Kiss and J. Strunk. "Unsupervised Multilingual Sentence Boundary Detection". In: *Computational Linguistics* 32.4 (2006), pp. 485–525. DOI: 10.1162/coli.2006.32.4.485. URL: https://aclanthology.org/J06-4003/.

[82] T. Kiss and J. Strunk. "Unsupervised Multilingual Sentence Boundary Detection". In: *Computational Linguistics* 32.4 (2006), pp. 485–525. DOI: 10.1162/coli.2006.32.4.485. URL: https://aclanthology.org/J06-4003/.

[83] A. Antonacopoulos, S. Pletschacher, D. Bridson, and C. Papadopoulos. "ICDAR 2009 Page Segmentation Competition". In: *2009 10th International Conference on Document Analysis and Recognition*. 2009, pp. 1370–1374. DOI: 10.1109/ICDAR.2009.275.

[84] M. Li, Y. Xu, L. Cui, S. Huang, F. Wei, Z. Li, and M. Zhou. *DocBank: A Benchmark Dataset for Document Layout Analysis*. 2020. arXiv: 2006.01038 [cs.CL]. URL: https://arxiv.org/abs/2006.01038.

[85] X. Zhong, J. Tang, and A. J. Yepes. "PubLayNet: largest dataset ever for document layout analysis". In: *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE. Sept. 2019, pp. 1015–1022. DOI: 10.1109/ICDAR.2019.00166.

[86] K. Benkirane. *publaynet-mini*. https : / / huggingface . co / datasets / kenza - ily / publaynet-mini. 2029.

[87] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye. *Object Detection in 20 Years: A Survey*. 2023. arXiv: 1905.05055 [cs.CV]. URL: https://arxiv.org/abs/1905.05055.

[88] A. J. Yepes, X. Zhong, and D. Burdick. *ICDAR 2021 Competition on Scientific Literature Parsing*. 2021. arXiv: 2106.14616 [cs.IR]. URL: https://arxiv.org/abs/2106.14616.

[89] K. Oksuz, B. C. Cam, S. Kalkan, and E. Akbas. *One Metric to Measure them All: Localisation Recall Precision (LRP) for Evaluating Visual Detection Tasks*. 2021. arXiv: 2011.10772 [cs.CV]. URL: https://arxiv.org/abs/2011.10772.

[90] I. Karmanov, A. S. Deshmukh, L. Voegtle, P. Fischer, K. Chumachenko, T. Roman, J. Seppänen, J. Parmar, J. Jennings, A. Tao, et al. "\'Eclair–Extracting Content and Layout with Integrated Reading Order for Documents". In: *arXiv preprint arXiv:2502.04223* (2025).

[91] R. Padilla, W. L. Passos, T. L. B. Dias, S. L. Netto, and E. A. B. da Silva. "A Comparative Analysis of Object Detection Metrics with a Companion Open-Source Toolkit". In: *Electronics* 10.3 (2021). ISSN: 2079-9292. DOI: 10 . 3390 / electronics10030279. URL: https://www.mdpi.com/2079-9292/10/3/279.

[92] R. Kaur and S. Singh. "A comprehensive review of object detection with deep learning". In: *Digital Signal Processing* 132 (2023), p. 103812. ISSN: 1051-2004. DOI: https://doi. org/10.1016/j.dsp.2022.103812. URL: https://www.sciencedirect.com/science/ article/pii/S1051200422004298.

[93] Z. C. Lipton, C. Elkan, and B. Narayanaswamy. *Thresholding Classifiers to Maximize F1 Score*. 2014. arXiv: 1402.1892 [stat.ML]. URL: https://arxiv.org/abs/1402.1892.

[94] A. Avetisyan, C. Xie, H. Howard-Jenkins, T.-Y. Yang, S. Aroudj, S. Patra, F. Zhang, D. Frost, L. Holland, C. Orme, et al. "Scenescript: Reconstructing scenes with an autoregressive structured language model". In: *European Conference on Computer Vision*. Springer. 2024, pp. 247–263.

[95] M. C. Hinojosa Lee, J. Braet, and J. Springael. "Performance Metrics for Multilabel Emotion Classification: Comparing Micro, Macro, and Weighted F1-Scores". In: *Applied Sciences* 14.21 (2024). ISSN: 2076-3417. DOI: 10.3390/app14219863. URL: https://www. mdpi.com/2076-3417/14/21/9863.

[96] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. "Microsoft coco: Common objects in context". In: *European conference on computer vision*. Springer. 2014, pp. 740–755.

[97] MiXaiLL76. "Faster-COCO-Eval: Faster and Enhanced COCO Evaluation Library". In: (2024).

[98] C. Cui, T. Sun, S. Liang, T. Gao, Z. Zhang, J. Liu, X. Wang, C. Zhou, H. Liu, M. Lin, Y. Zhang, Y. Zhang, H. Zheng, J. Zhang, J. Zhang, Y. Liu, D. Yu, and Y. Ma. *PaddleOCR-VL: Boosting Multilingual Document Parsing via a 0.9B Ultra-Compact Vision-Language Model*. 2025. arXiv: 2510.14528 [cs.CV]. URL: https://arxiv.org/abs/2510.14528.

[99] OpenAI, : A. Hurst, et al. *GPT-4o System Card*. 2024. arXiv: 2410.21276 [cs.CL]. URL: https://arxiv.org/abs/2410.21276.

[100] V. I. Levenshtein. "Binary Codes Capable of Correcting Deletions, Insertions and Reversals". In: *Soviet Physics Doklady* 10 (Feb. 1966), p. 707.

[101] L. Yujian and L. Bo. "A Normalized Levenshtein Distance Metric". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29.6 (2007), pp. 1091–1095. DOI: 10.1109/TPAMI.2007.1078.

[102] M. Pawlik and N. Augsten. "Tree edit distance: Robust and memory-efficient". In: *Information Systems* 56 (2016), pp. 157–173. ISSN: 0306-4379. DOI: https://doi.org/10.1016/j.is.2015.08.004. URL: https://www.sciencedirect.com/science/article/pii/S0306437915001611.

[103] X. Zhong, E. ShafieiBavani, and A. J. Yepes. *Image-based table recognition: data, model, and evaluation*. 2020. arXiv: 1911.10683 [cs.CV]. URL: https://arxiv.org/abs/1911.10683.

[104] S. Es, J. James, L. Espinosa-Anke, and S. Schockaert. *Ragas: Automated Evaluation of Retrieval Augmented Generation*. 2025. arXiv: 2309.15217 [cs.CL]. URL: https://arxiv.org/abs/2309.15217.

[105] F. Shi, X. Chen, K. Misra, N. Scales, D. Dohan, E. Chi, N. Schärli, and D. Zhou. *Large Language Models Can Be Easily Distracted by Irrelevant Context*. 2023. arXiv: 2302.00093 [cs.CL]. URL: https://arxiv.org/abs/2302.00093.

[106] J. Gallier. *Discrete mathematics*. Springer Science & Business Media, 2011. Chap. Strings, Multisets, Indexed Families, p. 217.

[107] S. R. Bhat, M. Rudat, J. Spiekermann, and N. Flores-Herr. *Rethinking Chunk Size For Long-Document Retrieval: A Multi-Dataset Analysis*. 2025. arXiv: 2505.21700 [cs.IR]. URL: https://arxiv.org/abs/2505.21700.

[108] LlamaIndex. *Evaluating the Ideal Chunk Size for a RAG System using LlamaIndex*. URL: https://www.llamaindex.ai/blog/evaluating-the-ideal-chunk-size-for-a-rag-system-using-llamaindex-6207e5d3fec5 (visited on 02/13/2026).

[109] A. Hannun, J. Digani, A. Katharopoulos, and R. Collobert. *MLX: Efficient and flexible machine learning on Apple silicon*. Version 0.0. 2023. URL: https://github.com/ml-explore.

[110] OpenAI Developers. *text-embedding-3-small: Small embedding model*. URL: https://developers.openai.com/api/docs/models/text-embedding-3-small (visited on 02/16/2026).

[111] Chroma. *Chroma: Open-source search and retrieval database for AI applications*. URL: https://github.com/chroma-core/chroma (visited on 02/15/2026).

[112] Z. Khan and Y. Fu. *Consistency and Uncertainty: Identifying Unreliable Responses From Black-Box Vision-Language Models for Selective Visual Question Answering*. 2024. arXiv: 2404.10193 [cs.CV]. URL: https://arxiv.org/abs/2404.10193.

[113] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. "Bleu: a method for automatic evaluation of machine translation". In: *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 2002, pp. 311–318.

[114] C.-Y. Lin. "Rouge: A package for automatic evaluation of summaries". In: *Text summarization branches out*. 2004, pp. 74–81.

[115] S. Banerjee and A. Lavie. "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments". In: *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*. 2005, pp. 65–72.

[116] F. Cuconasu, G. Trappolini, F. Siciliano, S. Filice, C. Campagnano, Y. Maarek, N. Tonellotto, and F. Silvestri. "The Power of Noise: Redefining Retrieval for RAG Systems". In: *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR 2024. ACM, July 2024, pp. 719–729. DOI: 10.1145/3626772.3657834. URL: http://dx.doi.org/10.1145/3626772.3657834.