Document Intelligence in the Era of Large Language Models: A Survey

Anonymous ACL submission

Abstract

Document AI (DAI) has emerged as a vital application area, and is significantly transformed by the advent of large language models (LLMs). While earlier approaches relied on encoder-decoder architectures, decoder-only LLMs have revolutionized DAI, bringing remarkable advancements in understanding and generation. This survey provides a comprehensive overview of DAI's evolution, highlighting current research attempts and future prospects of LLMs in this field. We explore key advancements and challenges in multimodal, multilingual, and retrieval-augmented DAI, while also suggesting future research directions, including agent-based approaches and document-specific foundation models. This paper aims to provide a structured analysis of the state-of-the-art in DAI and its implications for both academic and practical applications.

1 Introduction

Documents serve as diverse formats for conveying information, playing crucial roles in both research and industry (Stage and Manning, 2003). Document AI (DAI) leverages natural language processing and computer vision techniques to automate document-related tasks in two categories: understanding (Pfitzmann et al., 2022; Mathew et al., 2021) and generation (Wiseman et al., 2017; Zhang et al., 2018). However, traditional rule-based (Bourgeois et al., 1992) and learning-based (Marinai et al., 2005) approaches often prove cost-prohibitive and time-consuming, motivating the shift towards language models for more efficient and scalable solutions.

Among these models, *encoder-only* models (Hong et al., 2022; Huang et al., 2022; Luo et al., 2023) are widely used to capture nuanced, document-specific information, while *encoder-decoder* architectures excel in generation tasks involving variable-length and multimodal inputs and outputs (Tanaka et al., 2021; Tang et al., 2023), though they remain constrained by their training data. Recent advancements in *decoder-only* Large Language Models (LLMs) have spurred the development of document-focused LLMs trained on diverse multimodal datasets, such as image segmentation (Hu et al., 2024a), layout (Wang et al., 2024a; Luo et al., 2024), and geometric information (Luo et al., 2023; Lamott et al., 2024), demonstrating promising performance in both understanding and generating document content. Despite recent successes, LLMs struggle to interpret documents accurately (Liu et al., 2024a), often relying on Optical Character Recognition (OCR) engines or neglecting the rich textual information within documents. Such limitations hinder the learning of a unified representation across multiple document-specific modalities. Moreover, there remain pervasive challenges in multilingualism and linguistic diversity (Joshi et al., 2020), as data scarcity and the lack of high-quality training data restrict the performance of multilingual LLMs. While the effectiveness of cross-lingual generalization has been explored for translation-equivalent tasks (Zhang et al., 2023), DAI often requires complex contextual information, including document structure and language-specific knowledge (Zhao et al., 2024c; Xing et al., 2024), which can further limit LLM performance in multilingual document processing.

The development and application of LLMs in DAI have emerged as a key research area. Multilingual and multimodal capabilities are essential for enabling these models to effectively handle diverse, real-world scenarios (Xu et al., 2024b; Zhu et al., 2024b). Recent efforts have focused on integrating multilingualism and multimodality to enhance unified document representation learning, thereby improving the versatility and robustness of LLMs (Hu et al., 2024a). Given the rapid evolution of LLM-based DAI, there is an increasing need for a comprehensive overview of their capabilities in document processing. To address this need, we survey recent advancements in state-of-the-art (SoTA) multilingual and multimodal LLMs, highlighting their progress in DAI, the challenges they encounter, and future prospects.

In this in-depth survey paper, we systematically categorize relevant research on LLM-based DAI into five key tasks. Our contributions include:

- We introduce the detailed definition and trace the evolution of document intelligence, along with an overview of relevant benchmarks in Section 2.
- We comprehensively review recent advancements in integrating multimodal capabilities in LLMs, emphasizing their effectiveness on diverse DAI applications in Section 3.
- We systematically analyze the advancements in the multilingual capabilities (Section 4) and retrieval-augmented paradigm (Section 5), exploring their impact on improving context-aware document intelligence.
- We summarize key technical areas addressing current challenges in developing reliable document-specific foundation models, outlining actionable prospects for future developments in DAI.

2 Task Definitions and Benchmarks

Task Definitions Research in DAI has gained significant attention and can be broadly categorized into document-related *understanding* and *generation* tasks. We define understanding tasks in DAI as those aimed at extracting and analyzing information from existing documents, while generation tasks in DAI involve creating new content and responses based on given documents and instructions. In essence, understanding focuses on encoding information within documents, while generation emphasizes decoding and producing content and responses. Notably, these two categories often overlap. Multilingual and multi-document pose additional challenges to DAI. However, given their frequent occurrence in real-world applications, they are crucial for real-world DAI (Xu et al., 2021b; Hu et al., 2024b). Relevant research works in the LLM era will be discussed in Section 3-5, while early work and evolution are provided in Appendix A.

For the understanding tasks, Key Information Extraction (KIE) involves extracting essential information from unstructured or semi-structured documents (Abdallah et al., 2024b). This process typically comprises two components: OCR for converting images of text into machine-readable format, and named entity recognition (NER) for classifying entities into predefined categories (Li et al., 2022a). Document Layout Analysis (DLA) focuses on analyzing the spatial arrangement of documents to understand their structure and layout, which is crucial for various downstream applications (Bin-Makhashen and Mahmoud, 2020). For Document-level tasks, Document Sentiment Analysis (DSA) determines the sentiment expressed in documents, providing insights into the emotional tone of the context, while Document Classification (DC) categorizes documents based on their content.

In the generation category, Document Summarization (DS) aims to create concise summaries of documents (ge Yao et al., 2017; Ma et al., 2020). Document Content Generation (DCG) involves generating new document content based on existing materials, such as generating textual continuations or creating figures and tables derived from the provided documents (Biswas et al., 2024). Question Answering (QA) focuses on generating accurate natural language responses to a given question, based on the context of documents (Zhu et al., 2021; Mathew et al., 2021). However, QA can involve both understanding and generation tasks, *i.e.*, extractive and generative QA. Given the focus LLMs, we classify it primarily as a generation task.

Benchmarks and Datasets Numerous comprehensive benchmarks have been established to evaluate DAI. A summary of these benchmarks, detailing languages, document counts, modalities, tasks, and open-source availability, is presented in Appendix B.

3 Multimodal Document AI (DAI)

DAI has evolved significantly with the advent of multimodal AI, incorporating various information sources to comprehensively understand documents.

Textual modality remains the primary information for most documents, whether extracted through OCR from scanned images or directly accessed from digital formats like PDFs. Early approaches treat text as a logical sequence (Niyogi and Srihari, 1986), forming the foundation for diverse document processing tasks. *Visual modalities* contain rich signals, including fonts, pictures, handwriting, and background designs (Viana and Oliveira, 2017; Yi et al., 2017). These elements are essential for tasks like logo identification, signature verification, and analysis of non-textual layout components. Modern vision-language models have revolutionized document processing by integrating image features alongside text (Huang et al., 2022).

This integration enables end-to-end processing that transcends the limitations of traditional OCR-based methods. Moreover, *layout modalities* focus on the spatial arrangement of document content, including bounding box coordinates, column structures, and region segmentation. This information is crucial for understanding relationships between document segments, such as key-value pairs within form regions. Advanced approaches either embed spatial information directly into the token sequence (Wang et al., 2023a; He et al., 2023; Perot et al., 2024) or utilize separate encoders for bounding boxes (Wang et al., 2024a; Luo et al., 2024; Liao et al., 2024), enabling models jointly processing textual and spatial information. Additionally, many documents also incorporate additional elements, *e.g.*, tables (Herzig et al., 2020) and charts (Luo et al., 2021).

Building on these multimodal capabilities, the next sections explore understanding and generation tasks in DAI, highlighting how LLMs are reshaping multimodal DAI to meet real-world needs.

3.1 Multimodal Understanding in DAI

Multimodal understanding in DAI involves *analyzing*, *extracting*, and *classifying* document information by combining diverse modalities (Srihari et al., 1986; Taylor et al., 1994). Leveraging LLMs, this field has evolved through two primary methods: *prompt/instruction-based* and *unified encoding* approaches, each offering unique advantages in fusing textual, visual, and layout modalities.

Document Layout Analysis (DLA) focuses on detecting and classifying structural elements within documents, such as text blocks, headers, and tables, as well as understanding their relationships. DLA is also crucial as an auxiliary training objective to support other tasks like KIE and QA. Fan et al. (2024a) has benchmarked various LLMs (*e.g.*, GPT-3.5, GPT-4o [1], and LLaMA2 (Touvron et al., 2023)) on document element relation extraction, revealing that general-purpose LLMs underperform those featured models based on RoBERTa (Liu et al., 2019b). Meanwhile, In Context Learning (ICL) and fine-tuning are insufficient in enhancing LLMs' document hierarchy parsing capabilities (Xing et al., 2024; Luo et al., 2023). Simple *Prompt/Instruction-based approaches* leverage the capabilities of LLMs, directly integrating layout information into text prompts. To this end, layout-aware Chain-of-Thoughts (CoT) have demonstrated efficacy in enhancing spatial reasoning capabilities, during both pretraining and fine-tuning stages (Liao et al., 2024). LLM document layout understanding capabilities are improved through layout-aware prompting (Lamott et al., 2024) without requiring additional learning and fine-tuning with layout-specific instructions (Luo et al., 2024). On the other hand, *Unified Encoding approaches* can deeply integrate text, spatial coordinates, and image features, capturing layout patterns for document segmentation and classification through large-scale pertaining (Xu et al., 2021a; Huang et al., 2022). LayoutLLM (Luo et al., 2024) and DocLayLLM (Liao et al., 2024) adopt LayoutLMv3 (Huang et al., 2022) as the image and layout encoders, and train their parameters with Document Layout Analysis objectives. Inspired by human reading strategies, Nguyen et al. (2021) further propose to selectively focus on key document regions rather than processing all tokens equally. While, Wang et al. (2024a) incorporates a disentangled spatial attention mechanism and an infilling pretraining objective, effectively capturing relationships between various document fields.

Key Information Extraction (KIE) focuses on identifying and extracting specific elements from documents, such as form fields and key-value pairs. *Prompt-based approaches* have explored various methods of incorporating spatial information within prompts (Lamott et al., 2024; He et al., 2023) besides textual information, including bounding boxes, geometric formats and HTML markup. Similarly, Perot et al. (2024) propose to leverage horizontal and vertical coordinates to capture spatial information. Additionally, LayTextLLM (Lu et al., 2024) introduces "box tokens" to represent bounding boxes, aligning them with text using LoRA (Hu et al., 2022) for improved layout parsing. However, coordinate-as-token approaches significantly increase token length, leading to higher computational costs. To enhance spatial understanding, few-shot demonstrations (He et al., 2023) or fine-tuning (Perot et al., 2024) are often required, further exacerbating sequence length and GPU constraints. *Unified Encoding approach* is an efficient way with encoding multiple modalities into hidden vectors, allowing robust KIE from complex layout patterns (Appalaraju et al., 2024; Wang et al., 2024a). For example, InstructDoc (Tanaka et al., 2024) features an extra “Document-former” encoder that fuses OCR tokens with learnable tokens before forwarding them to the main LLM, enhancing field detection and entity extraction. Addition-

[1] https://chat.openai.com/chat

ally, visually guided generative text-layout pretraining offers a hierarchical approach that functions as a native OCR model while simultaneously modeling text and layout (Mao et al., 2024). Meanwhile, DoCo(Li et al., 2024c) utilizes contrastive learning to align fine-grained document-object features with holistic visual representations, addressing the challenge of feature collapse in text-rich scenarios.

Document Classification (DC) identifies the category, type, or domain of a document (*e.g.*, invoice vs. resume), often leveraging textual, visual, and structural cues. *Prompt/Instruction-Based approaches* have leveraged LLMs to classify documents based on natural language instructions, offering flexible adaptation to unseen document types without retraining. Recent research on LLM-based methods has demonstrated the effectiveness of LLMs in extracting core document information (Liu and Healey, 2023; Hewapathirana et al., 2023), adapting for document classification by identifying features of each document type. Moreover, *Unified Encoding approaches* integrate multiple modalities (*e.g.*, OCR-based text, layout, and images) into a unified transformer model (Xu et al., 2021a; Huang et al., 2022). These approaches excel in tasks where document layout (*e.g.*, form structures) is a strong signal for classification (Wang et al., 2024a), improving classification accuracy and efficiency. In this context, Powalski et al. (2021) employs a decoder capable of integrating textual, visual, and layout-based information for document understanding, achieving SoTA performance on classification and retrieval tasks while simplifying the end-to-end processing pipeline. And LayoutXLM (Xu et al., 2021b) integrates text, layout, and image modalities to classify visually-enriched documents across different languages, demonstrating significant improvements in multilingual document contexts. By unifying multiple document-related modalities, the dependency on external OCR engines is mitigated. These OCR-free approaches can further reduce computational overhead and text-based bottlenecks (Kim et al., 2022) by directly processing document images using a Transformer model. By leveraging the strengths of prompt-based, unified encoding, these models are pushing the boundaries of document analysis and classification capabilities.

3.2 Multimodal Generation in DAI

Multimodal document generation refers to the creation of new information based on the document input provided, either in the form of *question-answers*, *summaries*, or *document elements*. With the advent of modern LLM-based techniques, they can be used as black-box tools for information generation via prompting to integrate OCR; or other methods as white-box models, embedding multimodal representations directly into the more controllable generation process.

Document Question Answering (QA) focuses on answering questions based on documents, involving complex table lookups, referencing figures, or extracting text from specific regions. Leveraging the efficacy of *Prompt/instruction-based approaches*, Ye et al. (2023) extends a vision-language LLM with document-specific instructions, utilizing strong image-text alignment learned during pretraining. It enables QA on scanned documents without explicit OCR. While Wang et al. (2023a) demonstrates that LLMs can effectively capture layout information through manufactured prompts using spaces and line breaks, highlighting the potential of simple yet effective prompt engineering techniques. Similarly, LATIN (Wang et al., 2023a) preserves layout cues by inserting strategic spaces and line breaks in text prompts, significantly improving few-shot QA performance. Moreover, *Unified Encoding approaches* integrate multiple modalities into a single and cohesive framework, supporting end-to-end QA without separate structure-parsing steps (Mathew et al., 2021; Chen et al., 2023c). Kim et al. (2023) employs a BART-style (Lewis et al., 2020a) encoder to process text and bounding box coordinates, coupled with an LLM decoder for answer generation. In contrast, UDOP (Tang et al., 2023) unifies vision, text, and layout features in a single Transformer, utilizing a generative mask-prediction framework to infill coherent answers conditioned on both textual and visual cues. Inspired by DocLLM (Wang et al., 2024a), enabling the model to locate and interpret relevant regions to address questions in documents with irregular or complex layouts. Layout-aware decoding strategies have demonstrated the potential of integrating spatial and visual information directly into the decoding process, enhancing the model's ability to understand and reason about document structure for accurate answer generation.

Document Summarization (DS) aims to create concise overviews while preserving essential content, where recent advancements have explored various pathways to achieve this goal. *Visually-Enriched approaches* leverage large-scale document-level representations (Zhang et al., 2019; Lee et al., 2022) to enhance summarization performance, capturing comprehensive semantic information. Models like PaLI-X (Chen et al., 2023c) demonstrate versatility by encoding entire pages, including text and images, to generate structured summaries. While image-text alignment techniques strengthen cross-modal coherence through pseudo image captions (Jiang et al., 2023), effectively integrating visual and textual information. Moreover, *Structure-Aware approaches* incorporate additional modalities beyond text-image integration to enrich document summaries, where Liu et al. (2022) integrates tabular data into financial and business reports, ensuring table-aware insights are effectively reflected in summaries. And layout-aware summarization utilizes document layout analysis to capture both physical and logical structures, maintaining the integrity of the original document's information flow. Besides the aforementioned approaches, strategies have been developed to address challenges while handling *longer or multiple documents*. Techniques focus on distilling key information (Xiao et al., 2022; Liu et al., 2024b), reducing redundancy across multiple resources. Chen et al. (2023b) have employed position interpolation, extending context windows to maintain coherence across large volumes of text within long documents. A monotone-submodular content selection approach has emerged to prioritize essential events across multiple sources (Kurisinkel and Chen, 2023). These developments reflect the ongoing evolution of producing more coherent, concise, and comprehensive overviews of complex materials.

Document Content Generation (DCG) focuses on automatically generating structured document layouts and textual content, emphasizing design and coherence in document composition. *Layout synthesis* seeks to dynamically structure documents based on input semantics. Advanced from heuristic rules to learning implicit relationships between textual and visual elements, deep generative models for content-aware graphic design modeling (Zheng et al., 2019) generate adaptable layouts aligned with content themes. LLM-based autoregressive document modeling goes beyond layout-specific generation, incorporating sequential dependencies between textual and structural elements. In this context, Biswas et al. (2024) jointly models document structure and content, synthesizing cohesive documents without relying solely on visual components. *Multimodal document synthesis* aims to generate textual, visual, and structural modalities with richer document representations. Methods like UDOP (Tang et al., 2023) leverage large-scale pretraining to unify vision, text, and layout, enabling comprehensive document understanding and generation across diverse domains. Similarly, StrucTexTv2 (Yu et al., 2023) introduces masked visual-textual pretraining, improving document structure modeling while eliminating OCR dependencies, achieving robust document processing. Meanwhile, some methods tailored for specialized document elements, such as ChartLlama (Han et al., 2023), focus on improving chart understanding and generation by leveraging multimodal instruction tuning.

4 Multilingual Document AI

There are over 7,000 spoken languages worldwide (Joshi et al., 2020) and a significant portion of online content is written in languages other than English (Xu et al., 2021b). This poses challenges to support multilingual DAI. Models trained on monolingual data unsurprisingly struggle with multilingual tasks due to their limited capacity to capture cross-linguistic nuances and cultural intricacies. In contrast, LLMs, trained on vast multilingual data, exhibit remarkable capabilities in DAI tasks (Bandarkar et al., 2024; Lee et al., 2022; Scao et al., 2022; García-Ferrero et al., 2024), though challenges still remain.

4.1 Multilingual Prompt-Based Approaches

Many studies examine LLMs' zero-shot prompting capabilities across languages, revealing that they often outperform fine-tuned small models in KIE and DSA. However, their superiority is not consistently observed in other tasks (Abdallah et al., 2023; Bhat and Varma, 2023). Factors such as prompt quality and supplementary instructions may limit performance. Lai et al. (2023) find ChatGPT underperforms fine-tuned models on multilingual tasks but excels when prompted in English, a trend also seen in (Deshpande et al., 2024; Chen et al., 2024b). Cross-lingual thought prompts further explore self-translating instructions into English (Huang et al., 2023; Etxaniz et al., 2024).

ICL with non-English examples boosts KIE and DC performance (Abdallah et al., 2023), though the effectiveness is inconsistent (Engländer et al., 2024). To address this, strategies such as incorrect examples (Mo et al., 2024), multilingual semantic similarity (Tanwar et al., 2023), and word-level code-switching (Shankar et al., 2024) have been employed. Other prompt-based approaches include soft prompt tuning for cross-lingual relation extraction (Hsu et al., 2023), automated prompt construction from relation triples (Chen et al., 2022), and multi-turn QA for zero-shot KIE (Wei et al., 2024). Despite these efforts, challenges persist in multilingual DAI, such as prompt sensitivity and representation gaps, which are constrained by LLMs' inherent multilingual capabilities.

4.2 Multilingual Training Strategies

English-centric LLMs often struggle with capturing cultural and linguistic nuances, leading to limitations in DAI (Hershcovich et al., 2022; Navigli et al., 2023). To address this, PersianLLaMA (Abbasi et al., 2023) and JASMINE (Nagoudi et al., 2023) incorporate curated document collections for better alignment, while ArabianGPT (Koubaa et al., 2024) improves Arabic morphological processing through an optimized tokenizer. Additionally, preference-tuning techniques including reinforcement learning with AI feedback (RLAIF) (Lee et al., 2024) and direct preference optimization (DPO) (Rafailov et al., 2023) have been applied to align on cultural and linguistic preferences (Huang et al., 2024a; Jinnai, 2024). Lai et al. (2024) propose to translate English instructions and inputs into multiple target languages and align with human preference, showing promising performance on knowledge-intensive DAI tasks.

On the other hand, training on large-scale multilingual documents may introduce language interference, *i.e.*, the curse of multilinguality (Chang et al., 2024). To address this, multi-stage training paradigms are adopted, selectively fine-tuning different model components at various stages (Li et al., 2024b; Wang et al., 2023c), providing more accurate KIE capabilities. Other approaches (Zhang et al., 2024e; Kojima et al., 2024) apply language-specific neurons (Zhao et al., 2024a) to train with each monolingual data while preserving high-resource languages performance. Also, parameter-efficient techniques such as LoRA and adapters (Li et al., 2024b; Whitehouse et al., 2024; Zhao et al., 2024b) enable language-sensitive adaptation.

Another core challenge in multilingual DAI is achieving effective cross-lingual alignment. Nooralahzadeh and Sennrich (2023) improves alignment through similarity-based loss and synthetic code-mixing. PLUG (Zhang et al., 2024g) enhances linguistic structure alignment via pivot-language instructions. Contrastive learning, particularly effective for retrieval tasks (Wang et al., 2022b; Tan et al., 2023), improves alignment by clustering positive examples while separating negatives. Another approach leverages external parallel data, aligning text at the sentence (Zhu et al., 2023; Heffernan et al., 2022) or word level (Zhao et al., 2021; Chi et al., 2021b). Additionally, data augmentation strengthens alignment by generating synthetic data via translation (Zhang et al., 2024c) or self-distillation (Zhang et al., 2024f). Joshi et al. (2024) propose to use balanced synthetic corpora to continue pretraining LLMs.

These advancements underscore the potential of multilingual LLMs in DAI, though challenges in representation and effective cross-lingual generalization remain open research directions.

5 Retrieval-Augmented Paradigm

The retrieval-augmented paradigm has been widely studied in LLM research and industry. Many studies (Gao et al., 2023; Ram et al., 2023; Fan et al., 2024b; Hui et al., 2024) have shown that retrieving reliable external knowledge can mitigate challenges associated with outdated training data and limited domain expertise. Documents, encompassing multiple modalities like text, tables, and images, can be essentially integrated into this paradigm to support DAI (Zhao et al., 2023; Soman and Roychowdhury, 2024). For example, in business document applications, retrieval-augmented generation (RAG) is integrated into a document intelligence service, leveraging semantic chunking and layout information (Laujan, 2024). Similarly, it has been utilized to provide document grounding capabilities, allowing users to leverage extra documents for QA (Sebastienb, 2024).

5.1 Text-Based Retrieval Augmentation

Text, as the core component of documents, holds a wealth of information. Text-based retrieval augmentation can provide LLMs with precise and reliable contextual information, supporting them in understanding and generating tasks. Guu et al. (2020) introduce a pretrained retrieval-augmented language model (REALM). RAG is proposed around the same time, combining a pretrained model with a non-parametric memory storing a dense vector index of the external database (Lewis et al., 2020b). Subsequently, many studies have been conducted to enhance the performance and robustness of document-related tasks (Borgeaud et al., 2022; Izacard et al., 2023; Ram et al., 2023; Yu et al., 2024b; Yan et al., 2024; Zhang et al., 2024d).

For *KIE*, specifically event argument extraction, Li et al. (2021) propose a conditional model based on a predefined template, designed to handle missing arguments and enable cross-sentence inferences. Retrieve-and-Sample (Ren et al., 2023) is proposed to address the challenge of similar inputs leading to inconsistent outputs, which utilizes a hybrid RAG that samples pseudo demonstrations. DS and QA also benefit from RAG, *e.g.*, Edge et al. (2024) propose Graph RAG for query-focused summarization and DR-RAG (Hei et al., 2024) is introduced for QA by evaluating the dynamic relevance between query and documents.

For *long-context* DAI tasks, Xu et al. (2024a) find that using RAG can achieve comparable performance to fine-tuned LLM while taking much less computation. To improve long-context document retrieval, RAPTOR (Sarthi et al., 2024) is introduced to integrate information from a tree structure with differing levels of summarization to improve holistic document understanding. Dong et al. (2024) introduce a document-graph-based re-ranker to improve RAG. Besides, to evaluate the performance of long-document retrieval, Laban et al. (2024) introduce a synthetic testbed derived from summarization tasks and demonstrated that it plays a central role in assessing RAG effectiveness in long-context scenarios. Despite these efforts, long-context document retrieval remains a challenge, particularly in understanding how retrieval augmentation affects LLMs and how retrieved documents can be effectively incorporated into LLMs (Chen et al., 2023a).

5.2 Multimodal Retrieval Augmentation

Multimodal retrieval augmentation goes beyond using only text information, incorporating images, charts, and tables as retrieval targets (Zhao et al., 2023). By interpreting and leveraging varied data modalities, it aims to boost the performance of LLMs in downstream tasks (Fan et al., 2024b).

Heuristic approaches use different parsing tools for various modalities to obtain corresponding embeddings from given documents. RA-VQA (Lin and Byrne, 2022) integrate OCR, object detection, and image captioning to convert target images into textual data for subsequent retrieval. Meanwhile, tables in documents often appear in a semi-structured format, which requires table detection and parsing techniques (Ma et al., 2023a; Lin, 2024). For structured tabular data, T-RAG (Pan et al., 2022) utilizes vector index jointly with BART (Lewis et al., 2020a) to address QA. By leveraging table-to-table relevance and knowledge graphs to represent relationships between structured data, Chen et al. (2024a) and Sepasdar et al. (2024) enhance structured document retrieval.

Recently, advancements in vision-language models (VLMs) (OpenAI, 2023; Liu et al., 2023; Wang et al., 2024c) have enabled *end-to-end multimodal retrieval*, which reduces information loss and minimizes error accumulation. DSE (Ma et al., 2024a) is introduced to capture all information within a document by using screenshots and leveraging VLMs to encode these into retrieval representations. Similarly, ColPali (Faysse et al., 2024) utilizes ColBERT (Khattab and Zaharia, 2020) representations to index page-level documents, while VisRAG (Yu et al., 2024a), built on MiniCPM-V (Yao et al., 2024), supports page concatenation to improve retrieval performance. For visual document QA, VisDoMRAG (Suri et al., 2024) incorporates evidence curation and CoT to facilitate textual and visual RAGs, where the outputs are aligned to ensure consistency. Despite the significant development in multimodal retrieval, multimodal foundation models, effective embedding and representation methods, and comprehensive evaluation approaches remain open research directions.

6 Future Directions

AI agents are intelligent systems that exhibit a range of cognitive capabilities, including perception, learning, memory, planning, decision-making, action execution, and collaboration with other agents or humans (Huang, 2024; Huang et al., 2024b). Recent research has integrated reasoning and action within language models to tackle complex decision-making tasks (Yao et al., 2023), while advancements in learning from mistakes (Shinn et al., 2023) have further enhanced the performance of LLM agents. These developments have motivated applications across various fields, such as healthcare (Li et al., 2024a; Mehandru et al., 2024) software (Zhang et al., 2024b; Qian et al., 2024; Tang et al., 2024), and finance (Li et al., 2024d; Yang et al., 2024), demonstrating the transformative potential of AI agents. The rise of multilingual and multimodal LLMs (OpenAI, 2023) has positioned AI agents for document processing as a promising research area, offering numerous opportunities for innovation in handling complex document-related tasks.

Building on these insights, Document LLM agents (DocAgents) hold great potential for revolutionizing document processing. We define an autonomous DocAgent as an intelligent system designed to proficiently manage document understanding and generation tasks, achieving a level of expertise comparable to human specialists.

6.1 Collaborative DocAgent Framework

Documents typically combine rich textual representations, visual elements, and dynamic layouts (Ding et al., 2024; Zhang et al., 2024a). Managing these modalities often requires a specialized multi-agent framework, where effectively navigating these diverse modalities necessitates domain-expert agents. In this context, LLM-based multi-agent systems can collaborate by leveraging domain-specific knowledge, external tools, system feedback, and human inputs (Sprigler et al., 2024).

While existing frameworks have shown promising results in document simplification (Anonymous, 2024), they still encounter significant challenges in comprehending visually rich documents. Overcoming these obstacles requires integrating collaborative image-text retrieval capabilities and enhanced representations for complex document layouts, enabling a more comprehensive and nuanced approach to address complicated document-related tasks. Recent research indicates that even advanced multimodal LLMs struggle with structured data tasks, such as table representations (Deng et al., 2024). To advance DocAgent frameworks, future research endeavors should focus on: *(1) Enhanced Reasoning and Generalization:* Developing systems that produce documents tailored to specific input constraints, such as layout templates and natural language instructions. *(2) Complex Layout Handling:* Improving capabilities to manage dynamic, multimodal information within diverse document layouts. *(3) Efficient Information Retrieval:* Enabling global and local information searches with open-world interactions to capture relevant semantic nuances. Addressing these long-term challenges is key to developing reliable, explainable DocAgents for diverse document processing tasks.

6.2 DocAgent Foundation Model

While LLMs have emerged as general-purpose foundation models (Bommasani et al., 2021), they face two notable limitations: *(1) Domain-Specific Knowledge:* LLMs trained on broad web data often underperform on specialized tasks (Xie et al., 2024; Deng et al., 2024). *(2) Cross-Modal Alignment:* Current methods that utilize noisy web-text pairs and frozen encoder-decoder architectures can lead to misalignment across different modalities (Li et al., 2022b; Alayrac et al., 2022). These limitations can hinder the model's ability to integrate and process information from documents.

Recent efforts have introduced interactive foundation models that adapt to specific domains, such as robotics and healthcare (Durante et al., 2024). To address document-specific challenges, future work should prioritize: *(1) Cross-Modal Datasets:* Developing high-quality paired datasets that capture text, visual elements, and layout dynamics (Zhang et al., 2024a; Xing et al., 2024; Li et al., 2024e). *(2) Template-Based Information Extraction:* Leveraging question templates to extract key information with greater precision (Zmigrod et al., 2024). *(3) Enhanced Alignment Techniques:* Incorporating domain-specific question-answer agents to verify and refine multimodal data, ensuring consistency and accuracy across different modalities.

Developing a domain-aware DocAgent foundation model, supported by continuously evolving datasets, paves the way for an end-to-end solution for complex document processing tasks.

7 Conclusions

This survey provides a comprehensive review of recent advancements in Document AI (DAI), with a particular focus on LLM-based approaches. We have systematically outlined the evolving landscape of DAI, categorizing current tasks into two categories: understanding and generation. Our exploration reveals significant progress in integrating textual, layout, and visual modalities through prompt engineering or unified encoding strategies to handle DAI tasks. We also highlight the transformative impact of multilingual LLMs and retrieval-augmented methods on document processing capabilities, while also identifying remaining challenges in cross-lingual generalization, structural comprehension, and efficient multimodal learning. Last but not least, we discuss emerging trends, such as agent-based frameworks, as a promising path toward more robust and adaptable DAI systems.

Limitations

While we have strived to provide a comprehensive overview of recent advancements in the field of Document AI, certain limitations are inevitable.

First, our selection of works primarily focuses on cutting-edge LLM-based methods published in major conferences such as ACL, EMNLP, NAACL, NeurIPS, ICLR, and preprint repositories like arXiv over the last three years due to space constraints. While this focus reflects current trends, it may inadvertently lead to the exclusion of relevant contributions published in other venues or emerging close to or after the completion of this survey.

Second, while this survey emphasizes LLM-based approaches to align with recent developments in Document AI, non-LLM methods remain highly relevant and, in certain tasks and domains, may even be more effective. However, due to our focus, these approaches are not comprehensively covered, and their comparative advantages warrant further exploration.

Last but not least, the benchmarks discussed in this survey are among the most widely adopted in the Document AI community. While they provide a representative view of current evaluation practices, we acknowledge that other important benchmarks may not be fully captured. Future work could benefit from a broader investigation of evaluation metrics and benchmark datasets.

References

Mohammad Amin Abbasi, Arash Ghafouri, Mahdi Firouzmandi, Hassan Naderi, and Behrouz Minaei-Bidgoli. 2023. PersianLLaMA: Towards building first persian large language model. *CoRR*, abs/2312.15713.

Abdelrahman Abdallah, Mahmoud Abdalla, Mohamed Elkasaby, Yasser Elbendary, and Adam Jatowt. 2023. AMuRD: Annotated multilingual receipts dataset for cross-lingual key information extraction and classification. *CoRR*, abs/2309.09800.

Abdelrahman Abdallah, Mahmoud Abdalla, Mahmoud SalahEldin Kasem, Mohamed Mahmoud, Ibrahim Abdelhalim, Mohamed Elkasaby, Yasser Elbendary, and Adam Jatowt. 2024a. CORU: comprehensive post-OCR parsing and receipt understanding dataset. *CoRR*, abs/2406.04493.

Abdelrahman Abdallah, Daniel Eberharter, Zoe Pfister, and Adam Jatowt. 2024b. Transformers and language models in form understanding: A comprehensive review of scanned document analysis. *CoRR*, abs/2403.04080.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. 2022. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Abdullah Almutairi and Meshal Alfarhood. 2019. Instance segmentation of newspaper elements using mask R-CNN. In *18th IEEE International Conference On Machine Learning And Applications, ICMLA 2019, Boca Raton, FL, USA, December 16-19, 2019*, pages 1371–1375. IEEE.

Dhiraj Amin, Sharvari Govilkar, and Sagar Kulkarni. 2023. Question answering using deep learning in low resource Indian language Marathi. *CoRR*, abs/2309.15779.

Anonymous. 2024. Expertease: A multi-agent framework for grade-specific document simplification with large language models. In *Submitted to ACL Rolling Review - June 2024*. Under review.

Srikar Appalaraju, Peng Tang, Qi Dong, Nishant Sankaran, Yichu Zhou, and R. Manmatha. 2024. DocFormerv2: Local features for document understanding. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 709–718. AAAI Press.

Matheus Araújo, Júlio Cesar dos Reis, Adriano C. M. Pereira, and Fabrício Benevenuto. 2016. An evaluation of machine translation for multilingual sentence-level sentiment analysis. In *Proceedings of the 31st Annual ACM Symposium on Applied Computing, Pisa, Italy, April 4-8, 2016*, pages 1140–1145. ACM.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The Belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 749–775. Association for Computational Linguistics.

Savita Bhat and Vasudeva Varma. 2023. Large language models as annotators: A preliminary evaluation for annotating low-resource language content.

In *Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems, Eval4NLP 2023, Bali, Indonesia, November 1, 2023*, pages 100–107. Association for Computational Linguistics.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, Yuan-Fang Li, Yong-Bin Kang, and Rifat Shahriyar. 2023. CrossSum: Beyond English-centric cross-lingual summarization for 1, 500+ language pairs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 2541–2564. Association for Computational Linguistics.

Galal M. BinMakhashen and Sabri A. Mahmoud. 2020. Document layout analysis: A comprehensive survey. *ACM Comput. Surv.*, 52(6):109:1–109:36.

Sanket Biswas, Ayan Banerjee, Josep Lladós, and Umapada Pal. 2022. DocSegTr: An instance-level end-to-end document image segmentation transformer. *CoRR*, abs/2201.11438.

Sanket Biswas, Rajiv Jain, Vlad I. Morariu, Jiuxiang Gu, Puneet Mathur, Curtis Wigington, Tong Sun, and Josep Lladós. 2024. DocSynthv2: A practical autoregressive modeling for document generation. *CoRR*, abs/2406.08354.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ B. Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen Creel, Jared Quincy Davis, Dorottya Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, and et al. 2021. On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre. 2022. Improving language models by retrieving from trillions of tokens. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 2206–2240. PMLR.

Frank Le Bourgeois, Zbigniew Bublinski, and Hubert Emptoz. 1992. A fast and efficient method for extracting text paragraphs and graphics from unconstrained documents. In *11th IAPR International Conference on Pattern Recognition, ICPR 1992. Conference B: Pattern Recognition Methodology and Systems, The Hague, Netherlands, August 30-September 3, 1992*, pages 272–276. IEEE.

Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 335–336.

Ilias Chalkidis, Manos Fergadiotis, and Ion Androutsopoulos. 2021. MultiEURLEX - A multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 6974–6996. Association for Computational Linguistics.

Tyler A. Chang, Catherine Arnett, Zhuowen Tu, and Ben Bergen. 2024. When is multilinguality a curse? Language modeling for 250 high- and low-resource languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 4074–4096. Association for Computational Linguistics.

Hung-Ting Chen, Fangyuan Xu, Shane A. Arora, and Eunsol Choi. 2023a. Understanding retrieval augmentation for long-form question answering. *CoRR*, abs/2310.12150.

Peter Baile Chen, Yi Zhang, and Dan Roth. 2024a. Is table retrieval a solved problem? Exploring join-aware multi-table retrieval. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 2687–2699. Association for Computational Linguistics.

Shouyuan Chen, Sherman Wong, Liangjian Chen, and Yuandong Tian. 2023b. Extending context window of large language models via positional interpolation. *CoRR*, abs/2306.15595.

Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, Siamak Shakeri, Mostafa Dehghani, Daniel Salz, Mario Lucic, Michael Tschannen, Arsha Nagrani, Hexiang Hu, Mandar Joshi, Bo Pang, Ceslee Montgomery, Paulina Pietrzyk, Marvin Ritter, A. J. Piergiovanni, Matthias Minderer, Filip Pavetic, Austin Waters, Gang Li, Ibrahim Alabdulmohsin, Lucas Beyer, Julien Amelot, Kenton Lee, Andreas Peter Steiner, Yang Li, Daniel Keysers, Anurag Arnab, Yuanzhong Xu, Keran Rong, Alexander Kolesnikov, Mojtaba Seyedhosseini, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. 2023c. Pali-x: On scaling up a multilingual vision and language model. *CoRR*, abs/2305.18565.

Xingyu Chen, Zihan Zhao, Lu Chen, Jiabao Ji, Danyang Zhang, Ao Luo, Yuxuan Xiong, and Kai Yu. 2021. WebSRC: A dataset for web-based structural reading comprehension. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 4173–4185. Association for Computational Linguistics.

Yang Chen, Vedaant Shah, and Alan Ritter. 2024b. Translation and fusion improves zero-shot cross-lingual information extraction. *Preprint*, arXiv:2305.13582.

Yuxuan Chen, David Harbecke, and Leonhard Hennig. 2022. Multilingual relation classification via efficient and effective prompting. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 1059–1075. Association for Computational Linguistics.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021a. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 3576–3588. Association for Computational Linguistics.

Zewen Chi, Li Dong, Bo Zheng, Shaohan Huang, Xian-Ling Mao, Heyan Huang, and Furu Wei. 2021b. Improving pretrained cross-lingual language models via self-labeled word alignment. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 3418–3430. Association for Computational Linguistics.

Yew Ken Chia, Liying Cheng, Hou Pong Chan, Chaoqun Liu, Maojia Song, Sharifah Mahani Aljunied, Soujanya Poria, and Lidong Bing. 2024. M-Longdoc: A benchmark for multimodal super-long document understanding and a retrieval-aware tuning framework. *CoRR*, abs/2411.06176.

Naihao Deng, Zhenjie Sun, Ruiqi He, Aman Sikka, Yulong Chen, Lin Ma, Yue Zhang, and Rada Mihalcea. 2024. Tables as texts or images: Evaluating the table reasoning ability of LLMs and MLLMs. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 407–426. Association for Computational Linguistics.

Tejas Deshpande, Nidhi Kowtal, and Raviraj Joshi. 2024. Chain-of-translation prompting (CoTR): A novel prompting technique for low resource languages. *CoRR*, abs/2409.04512.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Pranjal Dhakal, Manish Munikar, and Bikram Dahal. 2019. One-shot template matching for automatic document data capture. In *2019 Artificial Intelligence for Transforming Business and Society (AITB)*, volume 1, pages 1–6. IEEE.

Yihao Ding, Jean Lee, and Soyeon Caren Han. 2024. Deep learning based visually rich document content understanding: A survey. *CoRR*, abs/2408.01287.

Jialin Dong, Bahare Fatemi, Bryan Perozzi, Lin F. Yang, and Anton Tsitsulin. 2024. Don't forget to connect! Improving RAG with graph-based reranking. *CoRR*, abs/2405.18414.

Li Dong and Mirella Lapata. 2016. Language to logical form with neural attention. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Kevin Duh, Akinori Fujino, and Masaaki Nagata. 2011. Is machine translation ripe for cross-lingual sentiment classification? In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA - Short Papers*, pages 429–433. The Association for Computer Linguistics.

Zane Durante, Bidipta Sarkar, Ran Gong, Rohan Taori, Yusuke Noda, Paul Tang, Ehsan Adeli, Shrinidhi Kowshika Lakshmikanth, Kevin A. Schulman, Arnold Milstein, Demetri Terzopoulos, Ade Famoti, Noboru Kuno, Ashley J. Llorens, Hoi Vo, Katsushi Ikeuchi, Li Fei-Fei, Jianfeng Gao, Naoki Wake, and Qiuyuan Huang. 2024. An interactive agent foundation model. *CoRR*, abs/2402.05929.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. 2024. From local to global: A graph RAG approach to query-focused summarization. *CoRR*, abs/2404.16130.

Harold P Edmundson. 1969. New methods in automatic extracting. *Journal of the ACM (JACM)*, 16(2):264–285.

Leon Engländer, Hannah Sterz, Clifton Poth, Jonas Pfeiffer, Ilia Kuznetsov, and Iryna Gurevych. 2024. M2QA: multi-domain multilingual question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 6283–6305. Association for Computational Linguistics.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lopez de Lacalle, and Mikel Artetxe. 2024. Do multilingual language models think better in English? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Short Papers, NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 550–564. Association for Computational Linguistics.

Shengda Fan, Yanting Wang, Shasha Mo, and Jianwei Niu. 2024a. LogicST: A logical self-training framework for document-level relation extraction with incomplete annotations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 5496–5510. Association for Computational Linguistics.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024b. A survey on RAG meeting LLMs: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2024, Barcelona, Spain, August 25-29, 2024*, pages 6491–6501. ACM.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2024. ColPali: Efficient document retrieval with vision language models. *CoRR*, abs/2407.01449.

Matthieu Futeral, Armel Zebaze, Pedro Ortiz Suarez, Julien Abadji, Rémi Lacroix, Cordelia Schmid, Rachel Bawden, and Benoît Sagot. 2024. mOSCAR: A large-scale multilingual and multimodal document-level corpus. *CoRR*, abs/2406.08707.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *CoRR*, abs/2312.10997.

Iker García-Ferrero, Rodrigo Agerri, Aitziber Atutxa Salazar, Elena Cabrio, Iker de la Iglesia, Alberto Lavelli, Bernardo Magnini, Benjamin Molinet, Johanna Ramirez-Romero, German Rigau, Jose Maria Villa-Gonzalez, Serena Villata, and Andrea Zaninello. 2024. Medical mT5: An open-source multilingual text-to-text LLM for the medical domain. *CoRR*, abs/2404.07613.

Lukasz Garncarek, Rafal Powalski, Tomasz Stanislawek, Bartosz Topolski, Piotr Halama, Michal Turski, and Filip Gralinski. 2021. LAMBERT: layout-aware language modeling for information extraction. In *16th International Conference on Document Analysis and Recognition, ICDAR 2021, Lausanne, Switzerland, September 5-10, 2021, Proceedings, Part I*, volume 12821 of *Lecture Notes in Computer Science*, pages 532–547. Springer.

Jin ge Yao, Xiaojun Wan, and Jianguo Xiao. 2017. Recent advances in document summarization. *Knowledge and Information Systems*, 53:297–336.

Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 580–587. IEEE Computer Society.

Albert Gu and Tri Dao. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *CoRR*, abs/2312.00752.

Zhangxuan Gu, Changhua Meng, Ke Wang, Jun Lan, Weiqiang Wang, Ming Gu, and Liqing Zhang. 2022. XYLayoutLM: Towards layout-aware multimodal networks for visually-rich document understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 4573–4582. IEEE.

Poonam Gupta and Vishal Gupta. 2012. A survey of text question answering techniques. *International Journal of Computer Applications*, 53(4).

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. REALM: retrieval-augmented language model pre-training. *CoRR*, abs/2002.08909.

Guangzeng Han, Jack Tsao, and Xiaolei Huang. 2024. Length-aware multi-kernel transformer for long document classification. *arXiv preprint arXiv:2405.07052*.

Yucheng Han, Chi Zhang, Xin Chen, Xu Yang, Zhibin Wang, Gang Yu, Bin Fu, and Hanwang Zhang. 2023. ChartLlama: A multimodal LLM for chart understanding and generation. *CoRR*, abs/2311.16483.

Adam W. Harley, Alex Ufkes, and Konstantinos G. Derpanis. 2015. Evaluation of deep convolutional nets for document image classification and retrieval. In *13th International Conference on Document Analysis and Recognition, ICDAR 2015, Nancy, France, August 23-26, 2015*, pages 991–995. IEEE Computer Society.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Samin Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-Sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 4693–4703. Association for Computational Linguistics.

Jiabang He, Lei Wang, Yi Hu, Ning Liu, Hui Liu, Xing Xu, and Heng Tao Shen. 2023. ICL-D3IE: in-context learning with diverse demonstrations updating for document information extraction. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 19428–19437. IEEE.

Kevin Heffernan, Onur Çelebi, and Holger Schwenk. 2022. Bitext mining using distilled sentence representations for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 2101–2112. Association for Computational Linguistics.

Zijian Hei, Weiling Liu, Wenjie Ou, Juyi Qiao, Junming Jiao, Guowen Song, Ting Tian, and Yi Lin. 2024. DR-RAG: applying dynamic document relevance to retrieval-augmented generation for question-answering. *CoRR*, abs/2406.07348.

Daniel Hershcovich, Stella Frank, Heather C. Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 6997–7013. Association for Computational Linguistics.

Jonathan Herzig, Pawel Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Eisenschlos. 2020. TaPas: Weakly supervised table parsing via pre-training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4320–4333, Online. Association for Computational Linguistics.

Kushan Hewapathirana, Nisansa de Silva, and C. D. Athuraliya. 2023. Multi-document summarization: A comparative evaluation. In *17th IEEE International Conference on Industrial and Information Systems, ICIIS 2023, Peradeniya, Sri Lanka, August 25-26, 2023*, pages 19–24. IEEE.

Tsutomu Hirao, Yasuhisa Yoshida, Masaaki Nishino, Norihito Yasuda, and Masaaki Nagata. 2013. Single-document summarization as a tree knapsack problem. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1515–1520, Seattle, Washington, USA. Association for Computational Linguistics.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing A multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 6609–6625. International Committee on Computational Linguistics.

Teakgyu Hong, Donghyun Kim, Mingi Ji, Wonseok Hwang, Daehyun Nam, and Sungrae Park. 2022. BROS: A pre-trained language model focusing on text and layout for better key information extraction from documents. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 10767–10775. AAAI Press.

Chiaming Hsu, Changtong Zan, Liang Ding, Longyue Wang, Xiaoting Wang, Weifeng Liu, Fu Lin, and Wenbin Hu. 2023. Prompt-learning for cross-lingual relation extraction. In *International Joint Conference on Neural Networks, IJCNN 2023, Gold Coast, Australia, June 18-23, 2023*, pages 1–9. IEEE.

Anwen Hu, Haiyang Xu, Jiabo Ye, Ming Yan, Liang Zhang, Bo Zhang, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. 2024a. mPLUG-DocOwl 1.5: Unified structure learning for OCR-free document understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 3096–3120. Association for Computational Linguistics.

Anwen Hu, Haiyang Xu, Liang Zhang, Jiabo Ye, Ming Yan, Ji Zhang, Qin Jin, Fei Huang, and Jingren Zhou. 2024b. mPLUG-DocOwl2: High-resolution compressing for OCR-free multi-page document understanding. *CoRR*, abs/2409.03420.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Pengfei Hu, Zhenrong Zhang, Jiefeng Ma, Shuhang Liu, Jun Du, and Jianshu Zhang. 2024c. DocMamba: Efficient document pre-training with state space model. *CoRR*, abs/2409.11887.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 12365–12394. Association for Computational Linguistics.

Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Dingjie Song, Zhihong Chen, Mosen Alharthi, Bang An, Juncai He, Ziche Liu, Junying Chen, Jianquan Li, Benyou Wang, Lian Zhang, Ruoyu Sun, Xiang Wan, Haizhou Li, and Jinchao Xu. 2024a. AceGPT, localizing large language models in Arabic.

In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 8139–8163. Association for Computational Linguistics.

Qiuyuan Huang, Naoki Wake, Bidipta Sarkar, Zane Durante, Ran Gong, Rohan Taori, Yusuke Noda, Demetri Terzopoulos, Noboru Kuno, Ade Famoti, Ashley J. Llorens, John Langford, Hoi Vo, Li Fei-Fei, Katsushi Ikeuchi, and Jianfeng Gao. 2024b. Position paper: Agent AI towards a holistic intelligence. *CoRR*, abs/2403.00833.

Yu Huang. 2024. Levels of AI agents: from rules to large language models. *CoRR*, abs/2405.06643.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. LayoutLMv3: Pre-training for document AI with unified text and image masking. In *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, pages 4083–4091. ACM.

Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and C. V. Jawahar. 2019. ICDAR2019 competition on scanned receipt OCR and information extraction. In *2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, September 20-25, 2019*, pages 1516–1520. IEEE.

Yulong Hui, Yao Lu, and Huanchen Zhang. 2024. UDA: A benchmark suite for retrieval augmented generation in real-world document analysis. *CoRR*, abs/2406.15187.

Gautier Izacard, Patrick S. H. Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot learning with retrieval augmented language models. *J. Mach. Learn. Res.*, 24:251:1–251:43.

Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. 2019. FUNSD: A dataset for form understanding in noisy scanned documents. In *2nd International Workshop on Open Services and Tools for Document Analysis, OST@ICDAR 2019, Sydney, Australia, September 22-25, 2019*, pages 1–6. IEEE.

Chaoya Jiang, Rui Xie, Wei Ye, Jinan Sun, and Shikun Zhang. 2023. Exploiting pseudo image captions for multimodal summarization. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 161–175. Association for Computational Linguistics.

Yuu Jinnai. 2024. Does cross-cultural alignment change the commonsense morality of language models? *CoRR*, abs/2406.16316.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1601–1611. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 6282–6293. Association for Computational Linguistics.

Raviraj Joshi, Kanishk Singla, Anusha Kamath, Raunak Kalani, Rakesh Paul, Utkarsh Vaidya, Sanjay Singh Chauhan, Niranjan Wartikar, and Eileen Long. 2024. Adapting multilingual LLMs to low-resource languages using continued pre-training and synthetic corpus. *CoRR*, abs/2410.14815.

Anoop Raveendra Katti, Christian Reisswig, Cordula Guder, Sebastian Brarda, Steffen Bickel, Johannes Höhne, and Jean Baptiste Faddoul. 2018. Chargrid: Towards understanding 2D documents. *arXiv preprint arXiv:1809.08799*.

Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval, SIGIR 2020, Virtual Event, China, July 25-30, 2020*, pages 39–48. ACM.

Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. 2022. OCR-free document understanding transformer. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXVIII*, volume 13688 of *Lecture Notes in Computer Science*, pages 498–517. Springer.

Geewook Kim, Hodong Lee, Daehee Kim, Haeji Jung, Sanghee Park, Yoonsik Kim, Sangdoo Yun, Taeho Kil, Bado Lee, and Seunghyun Park. 2023. Visually-situated natural language understanding with contrastive reading model and frozen large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 11989–12010. Association for Computational Linguistics.

Stefan Klink and Thomas Kieninger. 2001. Rule-based document structure understanding with a fuzzy combination of layout and textual features. *Int. J. Document Anal. Recognit.*, 4(1):18–26.

Tomás Kociský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The NarrativeQA reading comprehension challenge. *Trans. Assoc. Comput. Linguistics*, 6:317–328.

Takeshi Kojima, Itsuki Okimura, Yusuke Iwasawa, Hitomi Yanaka, and Yutaka Matsuo. 2024. On the multilingual ability of decoder-based pre-trained language models: Finding and controlling language-specific neurons. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 6919–6971. Association for Computational Linguistics.

Anis Koubaa, Adel Ammar, Lahouari Ghouti, Omar Najar, and Serry Sibaee. 2024. ArabianGPT: Native Arabic GPT-based large language model. *CoRR*, abs/2402.15313.

Jianfeng Kuang, Wei Hua, Dingkang Liang, Mingkun Yang, Deqiang Jiang, Bo Ren, and Xiang Bai. 2023. Visual information extraction in the wild: Practical dataset and end-to-end solution. In *Document Analysis and Recognition - ICDAR 2023 - 17th International Conference, San José, CA, USA, August 21-26, 2023, Proceedings, Part VI*, volume 14192 of *Lecture Notes in Computer Science*, pages 36–53. Springer.

Jayant Kumar, Peng Ye, and David S. Doermann. 2014. Structural similarity for document image classification and retrieval. *Pattern Recognit. Lett.*, 43:119–126.

Litton J. Kurisinkel and Nancy F. Chen. 2023. LLM based multi-document summarization exploiting main-event biased monotone submodular content extraction. *CoRR*, abs/2310.03414.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Trans. Assoc. Comput. Linguistics*, 7:452–466.

Philippe Laban, Alexander R. Fabbri, Caiming Xiong, and Chien-Sheng Wu. 2024. Summary of a haystack: A challenge to long-context LLMs and RAG systems. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 9885–9903. Association for Computational Linguistics.

Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen R. McKeown. 2020. Wikilingua: A new benchmark dataset for multilingual abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 4034–4048. Association for Computational Linguistics.

Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 13171–13189. Association for Computational Linguistics.

Wen Lai, Mohsen Mesgar, and Alexander Fraser. 2024. LLMs beyond English: Scaling the multilingual capability of LLMs with cross-lingual feedback. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 8186–8213. Association for Computational Linguistics.

Marcel Lamott, Yves-Noel Weweler, Adrian Ulges, Faisal Shafait, Dirk Krechel, and Darko Obradovic. 2024. LAPDoc: Layout-aware prompting for documents. In *Document Analysis and Recognition - ICDAR 2024 - 18th International Conference, Athens, Greece, August 30 - September 4, 2024, Proceedings, Part IV*, volume 14807 of *Lecture Notes in Computer Science*, pages 142–159. Springer.

Jordy Van Landeghem, Rafal Powalski, Rubèn Tito, Dawid Jurkiewicz, Matthew B. Blaschko, Lukasz Borchmann, Mickaël Coustaty, Sien Moens, Michal Pietruszka, Bertrand Anckaert, Tomasz Stanislawek, Pawel Józiak, and Ernest Valveny. 2023. Document understanding dataset and evaluation (DUDE). In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 19471–19483. IEEE.

Laujan. 2024. Retrieval-augmented generation (RAG) with Azure AI document intelligence.

Chia-Hsuan Lee and Hung-yi Lee. 2019. Cross-lingual transfer learning for question answering. *CoRR*, abs/1907.06042.

Chia-Hsuan Lee, Aditya Siddhant, Viresh Ratnakar, and Melvin Johnson. 2022. DOCmT5: Document-level pretraining of multilingual language models. In *Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 425–437. Association for Computational Linguistics.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. 2024. RLAIF vs. RLHF: scaling reinforcement learning from human feedback with AI feedback. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Anton Leuski, Chin-Yew Lin, Liang Zhou, Ulrich Germann, Franz Josef Och, and Eduard H. Hovy. 2003. Cross-lingual c*st*rd: English access to Hindi information. *ACM Trans. Asian Lang. Inf. Process.*, 2(3):245–269.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020b. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Binxu Li, Tiankai Yan, Yuanting Pan, Zhe Xu, Jie Luo, Ruiyang Ji, Shilong Liu, Haoyu Dong, Zihao Lin, and Yixin Wang. 2024a. MMedAgent: Learning to use medical tools with multi-modal agent. *CoRR*, abs/2407.02483.

Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2022a. A survey on deep learning for named entity recognition. *IEEE Trans. Knowl. Data Eng.*, 34(1):50–70.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022b. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR.

Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, Ming Zhou, and Zhoujun Li. 2020a. TableBank: Table benchmark for image-based table detection and recognition. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 1918–1925. European Language Resources Association.

Minghao Li, Yiheng Xu, Lei Cui, Shaohan Huang, Furu Wei, Zhoujun Li, and Ming Zhou. 2020b. DocBank: A benchmark dataset for document layout analysis. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 949–960. International Committee on Computational Linguistics.

Sha Li, Heng Ji, and Jiawei Han. 2021. Document-level event argument extraction by conditional generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 894–908. Association for Computational Linguistics.

Tongliang Li, Zixiang Wang, Linzheng Chai, Jian Yang, Jiaqi Bai, Yuwei Yin, Jiaheng Liu, Hongcheng Guo, Liqun Yang, Hebboul Zine El Abidine, and Zhoujun Li. 2024b. mt4CrossOIE: Multi-stage tuning for cross-lingual open information extraction. *Expert Syst. Appl.*, 255:124760.

Xin Li, Yunfei Wu, Xinghua Jiang, Zhihao Guo, Mingming Gong, Haoyu Cao, Yinsong Liu, Deqiang Jiang, and Xing Sun. 2024c. Enhancing visual document understanding with contrastive learning in large visual-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 15546–15555. IEEE.

Yuan Li, Bingqiao Luo, Qian Wang, Nuo Chen, Xu Liu, and Bingsheng He. 2024d. A reflective LLM-based agent to guide zero-shot cryptocurrency trading. *CoRR*, abs/2407.09546.

Zichao Li, Shaojie He, Meng Liao, Xuanang Chen, Yaojie Lu, Hongyu Lin, Yanxiong Lu, Xianpei Han, and Le Sun. 2024e. Seg2act: Global context-aware action generation for document logical structuring. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 18077–18088. Association for Computational Linguistics.

Zihao Li, Yucheng Shi, Zirui Liu, Fan Yang, Ninghao Liu, and Mengnan Du. 2024f. Quantifying multilingual performance of large language models across languages. *CoRR*, abs/2404.11553.

Yunlong Liang, Fandong Meng, Chulun Zhou, Jinan Xu, Yufeng Chen, Jinsong Su, and Jie Zhou. 2022. A variational hierarchical model for neural cross-lingual summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2088–2099. Association for Computational Linguistics.

Wenhui Liao, Jiapeng Wang, Hongliang Li, Chengyu Wang, Jun Huang, and Lianwen Jin. 2024. DocLayLLM: An efficient and effective multi-modal extension of large language models for text-rich document understanding. *CoRR*, abs/2408.15045.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Demiao Lin. 2024. Revolutionizing retrieval-augmented generation with enhanced PDF structure recognition. *CoRR*, abs/2401.12599.

Guo-Shiang Lin, Jia-Cheng Tu, and Jen-Yung Lin. 2021. Keyword detection based on retinanet and transfer learning for personal information protection in document images. *Applied Sciences*, 11(20):9528.

Hui Lin and Jeff Bilmes. 2011. A class of submodular functions for document summarization. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 510–520.

Weizhe Lin and Bill Byrne. 2022. Retrieval augmented visual question answering with outside knowledge. *CoRR*, abs/2210.03809.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Linlin Liu, Bosheng Ding, Lidong Bing, Shafiq R. Joty, Luo Si, and Chunyan Miao. 2021. MulDA: A multilingual data augmentation framework for low-resource cross-lingual NER. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 5834–5846. Association for Computational Linguistics.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024a. Lost in the middle: How language models use long contexts. *Trans. Assoc. Comput. Linguistics*, 12:157–173.

Ran Liu, Ming Liu, Min Yu, Jianguo Jiang, Gang Li, Dan Zhang, Jingyuan Li, Xiang Meng, and Weiqing Huang. 2024b. GLIMMER: incorporating graph and lexical features in unsupervised multi-document summarization. In *ECAI 2024 - 27th European Conference on Artificial Intelligence, 19-24 October 2024, Santiago de Compostela, Spain - Including 13th Conference on Prestigious Applications of Intelligent Systems (PAIS 2024)*, volume 392 of *Frontiers in Artificial Intelligence and Applications*, pages 3709–3716. IOS Press.

Sengjie Liu and Christopher G. Healey. 2023. Abstractive summarization of large document collections using GPT. *CoRR*, abs/2310.05690.

Shuaiqi Liu, Jiannong Cao, Ruosong Yang, and Zhiyuan Wen. 2022. Long text and multi-table summarization: Dataset and method. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 1995–2010. Association for Computational Linguistics.

Xiaojing Liu, Feiyu Gao, Qiong Zhang, and Huasha Zhao. 2019a. Graph convolution for multimodal information extraction from visually rich documents. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 2 (Industry Papers)*, pages 32–39. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Antoine Louis, Vageesh Saxena, Gijs van Dijck, and Gerasimos Spanakis. 2024. ColBERT-XM: A modular multi-vector representation model for zero-shot multilingual information retrieval. *CoRR*, abs/2402.15059.

Jinghui Lu, Haiyang Yu, Yanjie Wang, Yongjie Ye, Jingqun Tang, Ziwei Yang, Binghong Wu, Qi Liu, Hao Feng, Han Wang, et al. 2024. A bounding box is worth one token: Interleaving layout and text in a large language model for document understanding. *arXiv preprint arXiv:2407.01976*.

Hans Peter Luhn. 1958. The automatic creation of literature abstracts. *IBM Journal of research and development*, 2(2):159–165.

Chuwei Luo, Changxu Cheng, Qi Zheng, and Cong Yao. 2023. GeoLayoutLM: Geometric pre-training for visual information extraction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 7092–7101. IEEE.

Chuwei Luo, Yufan Shen, Zhaoqing Zhu, Qi Zheng, Zhi Yu, and Cong Yao. 2024. LayoutLLM: Layout instruction tuning with large language models for document understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 15630–15640. IEEE.

Junyu Luo, Zekun Li, Jinpeng Wang, and Chin-Yew Lin. 2021. ChartOCR: Data extraction from charts images via a deep hybrid framework. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1917–1925.

Chixiang Ma, Weihong Lin, Lei Sun, and Qiang Huo. 2023a. Robust table detection and structure recognition from heterogeneous document images. *Pattern Recognit.*, 133:109006.

Congbo Ma, W. Zhang, Mingyu Guo, Hu Wang, and Quan Z. Sheng. 2020. Multi-document summarization via deep learning techniques: A survey. *ACM Computing Surveys*, 55:1 – 37.

Jiefeng Ma, Jun Du, Pengfei Hu, Zhenrong Zhang, Jianshu Zhang, Huihui Zhu, and Cong Liu. 2023b. HRDoc: Dataset and baseline method toward hierarchical reconstruction of document structures. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI*

2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023, pages 1870–1877. AAAI Press.

Xueguang Ma, Sheng-Chieh Lin, Minghan Li, Wenhu Chen, and Jimmy Lin. 2024a. Unifying multimodal retrieval via document screenshot embedding. *CoRR*, abs/2406.11251.

Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, Pan Zhang, Liangming Pan, Yu-Gang Jiang, Jiaqi Wang, Yixin Cao, and Aixin Sun. 2024b. MMLONGBENCH-DOC: benchmarking long-context document understanding with visualizations. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Ibrahim Souleiman Mahamoud, Mickaël Coustaty, Aurélie Joseph, Vincent Poulain D'Andecy, and Jean-Marc Ogier. 2024. CHIC: corporate document for visual question answering. In *Document Analysis and Recognition - ICDAR 2024 - 18th International Conference, Athens, Greece, August 30 - September 4, 2024, Proceedings, Part VI*, volume 14809 of *Lecture Notes in Computer Science*, pages 113–127. Springer.

Zhiming Mao, Haoli Bai, Lu Hou, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. 2024. Visually guided generative text-layout pre-training for document intelligence. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 4713–4730. Association for Computational Linguistics.

Simone Marinai, Marco Gori, and Giovanni Soda. 2005. Artificial neural networks for document analysis and recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(1):23–35.

Marko Markovic and Stevan Gostojic. 2020. Knowledge-based legal document assembly. *CoRR*, abs/2009.06611.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V. Jawahar. 2022. InfographicVQA. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, Waikoloa, HI, USA, January 3-8, 2022*, pages 2582–2591. IEEE.

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. 2021. DocVQA: A dataset for VQA on document images. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021*, pages 2199–2208. IEEE.

Ryan McDonald. 2007. A study of global inference algorithms in multi-document summarization. In *European Conference on Information Retrieval*, pages 557–564. Springer.

Nikita Mehandru, Brenda Y. Miao, Eduardo Rodriguez Almaraz, Madhumita Sushil, Atul J. Butte, and Ahmed M. Alaa. 2024. Evaluating large language models as agents in the clinic. *npj Digit. Medicine*, 7(1).

Tomas Mikolov. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 3781.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. OCR-VQA: visual question answering by reading text in images. In *2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, September 20-25, 2019*, pages 947–952. IEEE.

Ying Mo, Jiahao Liu, Jian Yang, Qifan Wang, Shun Zhang, Jingang Wang, and Zhoujun Li. 2024. C-ICL: contrastive in-context learning for information extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024*, pages 10099–10114. Association for Computational Linguistics.

Thomas Mueller, Francesco Piccinno, Massimo Nicosia, Peter Shaw, and Yasemin Altun. 2019. Answering conversational questions on structured data without logical forms. *arXiv preprint arXiv:1908.11787*.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, Nedjma Ousidhoum, David Ifeoluwa Adelani, Seid Muhie Yimam, Ibrahim Said Ahmad, Meriem Beloucif, Saif M. Mohammad, Sebastian Ruder, Oumaima Hourrane, Alípio Jorge, Pavel Brazdil, Felermino Dário Mário António Ali, Davis David, Salomey Osei, Bello Shehu Bello, Falalu Ibrahim Lawan, Tajuddeen Gwadabe, Samuel Rutunda, Tadesse Destaw Belay, Wendimu Baye Messelle, Hailu Beshada Balcha, Sisay Adugna Chala, Hagos Tesfahun Gebremichael, Bernard Opoku, and Stephen Arthur. 2023. AfriSenti: A Twitter sentiment analysis benchmark for African languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 13968–13981. Association for Computational Linguistics.

El Moatez Billah Nagoudi, Muhammad Abdul-Mageed, AbdelRahim A. Elmadany, Alcides Alcoba Inciarte, and Md Tawkat Islam Khondaker. 2023. JASMINE: Arabic GPT models for few-shot learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 16721–16744. Association for Computational Linguistics.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. SummaRuNNer: A recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 3075–3081. AAAI Press.

Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. Ranking sentences for extractive summarization with reinforcement learning. *arXiv preprint arXiv:1802.08636*.

Roberto Navigli, Simone Conia, and Björn Ross. 2023. Biases in large language models: Origins, inventory, and discussion. *ACM J. Data Inf. Qual.*, 15(2):10:1–10:21.

Khang Tan Tran Minh Nguyen. 2024. Uit-mlreceipts: A multilingual benchmark for detecting and recognizing key information in receipts. *REV Journal on Electronics and Communications*, 14(1).

Laura Nguyen, Thomas Scialom, Jacopo Staiano, and Benjamin Piwowarski. 2021. Skim-attention: Learning to focus via document layout. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 2413–2427. Association for Computational Linguistics.

Son Lam Nguyen Vu, Tien Dong Nguyen, and Van Hai Pham. 2024. A robust component-based template matching approach using document layout graph for extracting information. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pages 10–22. Springer.

Debashish Niyogi and Sargur N. Srihari. 1986. A rule-based system for document understanding. In *Proceedings of the 5th National Conference on Artificial Intelligence. Philadelphia, PA, USA, August 11-15, 1986. Volume 2: Engineering*, pages 789–793. Morgan Kaufmann.

Farhad Nooralahzadeh and Rico Sennrich. 2023. Improving the cross-lingual generalisation in visual question answering. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 13419–13427. AAAI Press.

Sofia Ares Oliveira, Benoit Seguin, and Frédéric Kaplan. 2018. dhSegment: A generic deep-learning approach for document segmentation. In *16th International Conference on Frontiers in Handwriting Recognition, ICFHR 2018, Niagara Falls, NY, USA, August 5-8, 2018*, pages 7–12. IEEE Computer Society.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Feifei Pan, Mustafa Canim, Michael R. Glass, Alfio Gliozzo, and James A. Hendler. 2022. End-to-end table question answering via retrieval-augmented generation. *CoRR*, abs/2203.16714.

Seunghyun Park, Seung Shin, Bado Lee, Junyeop Lee, Jaeheung Surh, Minjoon Seo, and Hwalsuk Lee. 2019. CORD: A consolidated receipt dataset for post-OCR parsing. In *Document Intelligence Workshop at Neural Information Processing Systems*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Vincent Perot, Kai Kang, Florian Luisier, Guolong Su, Xiaoyu Sun, Ramya Sree Boppana, Zilong Wang, Zifeng Wang, Jiaqi Mu, Hao Zhang, Chen-Yu Lee, and Nan Hua. 2024. LMDX: language model-based document information extraction and localization. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 15140–15168. Association for Computational Linguistics.

Birgit Pfitzmann, Christoph Auer, Michele Dolfi, Ahmed S. Nassar, and Peter W. J. Staar. 2022. DocLayNet: A large human-annotated dataset for document-layout analysis. *CoRR*, abs/2206.01062.

Rafal Powalski, Lukasz Borchmann, Dawid Jurkiewicz, Tomasz Dwojak, Michal Pietruszka, and Gabriela Palka. 2021. Going full-tilt boogie on document understanding with text-image-layout transformer. In *16th International Conference on Document Analysis and Recognition, ICDAR 2021, Lausanne, Switzerland, September 5-10, 2021, Proceedings, Part II*, volume 12822 of *Lecture Notes in Computer Science*, pages 732–747. Springer.

Peter Prettenhofer and Benno Stein. 2010. Cross-language text classification using structural correspondence learning. In *ACL 2010, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, July 11-16, 2010, Uppsala, Sweden*, pages 1118–1127. The Association for Computer Linguistics.

Le Qi, Shangwen Lv, Hongyu Li, Jing Liu, Yu Zhang, Qiaoqiao She, Hua Wu, Haifeng Wang, and Ting Liu. 2022. DuReader$_{\text{vis}}$: A Chinese dataset for open-domain document visual question answering. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 1338–1351. Association for Computational Linguistics.

Chen Qian, Yufan Dang, Jiahao Li, Wei Liu, Zihao Xie, Yifei Wang, Weize Chen, Cheng Yang, Xin Cong, Xiaoyin Che, Zhiyuan Liu, and Maosong Sun. 2024. Experiential co-learning of software-developing agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 5628–5640. Association for Computational Linguistics.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Trans. Assoc. Comput. Linguistics*, 11:1316–1331.

Himashi Rathnayake, Janani Sumanapala, Raveesha Rukshani, and Surangika Ranathunga. 2022. Adapter-based fine-tuning of pre-trained multilingual language models for code-mixed and code-switched text classification. *Knowl. Inf. Syst.*, 64(7):1937–1966.

Johannes Rausch, Octavio Martinez, Fabian Bissig, Ce Zhang, and Stefan Feuerriegel. 2021. Docparser: Hierarchical document structure parsing from renderings. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 4328–4338. AAAI Press.

Johannes Rausch, Gentiana Rashiti, Maxim Gusev, Ce Zhang, and Stefan Feuerriegel. 2023. DSG: an end-to-end document structure generator. In *IEEE International Conference on Data Mining, ICDM 2023, Shanghai, China, December 1-4, 2023*, pages 518–527. IEEE.

Yubing Ren, Yanan Cao, Ping Guo, Fang Fang, Wei Ma, and Zheng Lin. 2023. Retrieve-and-sample: Document-level event argument extraction via hybrid retrieval augmentation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 293–306. Association for Computational Linguistics.

Ranajit Saha, Ajoy Mondal, and C. V. Jawahar. 2019. Graphical object detection in document images. In *2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, September 20-25, 2019*, pages 51–58. IEEE.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D. Manning. 2024. RAPTOR: recursive abstractive processing for tree-organized retrieval. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100.

Holger Schwenk and Xian Li. 2018. A corpus for multilingual document classification in eight languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. European Language Resources Association (ELRA).

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. MLSUM: the multilingual summarization corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 8051–8067. Association for Computational Linguistics.

Sebastienb. 2024. Harness retrieval-augmented generation in Joule and Generative AI Hub.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.

Zahra Sepasdar, Sushant Gautam, Cise Midoglu, Michael A Riegler, and Pål Halvorsen. 2024. Enhancing structured-data retrieval with graphrag: Soccer data case study. *arXiv preprint arXiv:2409.17580*.

Bhavani Shankar, Preethi Jyothi, and Pushpak Bhattacharyya. 2024. In-context mixing (ICM): Code-mixed prompts for multilingual LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4162–4176, Bangkok, Thailand. Association for Computational Linguistics.

Huawen Shen, Gengluo Li, Jinwen Zhong, and Yu Zhou. 2024. LDP: Generalizing to multilingual visual information extraction by language decoupled pretraining. *arXiv e-prints*, pages arXiv–2412.

Xiaoyu Shen, Akari Asai, Bill Byrne, and Adrià de Gispert. 2023. xPQA: Cross-lingual product question answering across 12 languages. *CoRR*, abs/2305.09249.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Stepán Simsa, Milan Sulc, Michal Uricár, Yash Patel, Ahmed Hamdi, Matej Kocián, Matyás Skalický, Jirí Matas, Antoine Doucet, Mickaël Coustaty, and Dimosthenis Karatzas. 2023. DocILE benchmark for document information localization and extraction. In *Document Analysis and Recognition - ICDAR 2023 - 17th International Conference, San José, CA, USA, August 21-26, 2023, Proceedings, Part II*, volume 14188 of *Lecture Notes in Computer Science*, pages 147–166. Springer.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards VQA models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326.

Sumit Soman and Sujoy Roychowdhury. 2024. Observations on building RAG systems for technical documents. In *The Second Tiny Papers Track at ICLR 2024, Tiny Papers @ ICLR 2024, Vienna, Austria, May 11, 2024*. OpenReview.net.

Asher Sprigler, Alexander Drobek, Keagan Weinstock, Wendpanga Tapsoba, Gavin Childress, Andy Dao, and Lucas Gral. 2024. Synergistic simulations: Multi-agent problem solving with large language models. *CoRR*, abs/2409.13753.

Sargur N Srihari, Stephen W Lam, Venu Govindaraju, R Srihari, and Jonathan J Hull. 1986. Document image understanding. In *FJCC*, pages 87–95. Citeseer.

Frances K Stage and Kathleen Manning. 2003. Research in the college context. *New York: Brunner*.

Tomasz Stanislawek, Filip Gralinski, Anna Wróblewska, Dawid Lipinski, Agnieszka Kaliska, Paulina Rosalska, Bartosz Topolski, and Przemyslaw Biecek. 2021. Kleister: Key information extraction datasets involving long documents with complex layouts. In *16th International Conference on Document Analysis and Recognition, ICDAR 2021, Lausanne, Switzerland, September 5-10, 2021, Proceedings, Part I*, volume 12821 of *Lecture Notes in Computer Science*, pages 564–579. Springer.

Hongbin Sun, Zhanghui Kuang, Xiaoyu Yue, Chenhao Lin, and Wayne Zhang. 2021. Spatial dual-modality graph reasoning for key information extraction. *CoRR*, abs/2103.14470.

Manan Suri, Puneet Mathur, Franck Dernoncourt, Kanika Goswami, Ryan A. Rossi, and Dinesh Manocha. 2024. VisDoM: Multi-document QA with visually rich elements using multimodal retrieval-augmented generation. *CoRR*, abs/2412.10704.

Hiroya Takamura and Manabu Okumura. 2009. Text summarization model based on maximum coverage problem and its variant. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 781–789.

Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. 2021. MultiModalQA: Complex question answering over text, tables and images. *arXiv preprint arXiv:2104.06039*.

Weiting Tan, Kevin Heffernan, Holger Schwenk, and Philipp Koehn. 2023. Multilingual representation distillation with contrastive learning. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 1469–1482. Association for Computational Linguistics.

Ryota Tanaka, Taichi Iki, Kyosuke Nishida, Kuniko Saito, and Jun Suzuki. 2024. InstructDoc: A dataset for zero-shot generalization of visual document understanding with instructions. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 19071–19079. AAAI Press.

Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. 2023. SlideVQA: A dataset for document visual question answering on multiple images. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 13636–13645. AAAI Press.

Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. 2021. VisualMRC: Machine reading comprehension on document images. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13878–13888. AAAI Press.

Daniel Tang, Zhenghan Chen, Kisub Kim, Yewei Song, Haoye Tian, Saad Ezzini, Yongfeng Huang, and Jacques Klein Tegawende F Bissyande. 2024. Collaborative agents for software engineering. *arXiv preprint arXiv:2402.02172*.

Zineng Tang, Ziyi Yang, Guoxin Wang, Yuwei Fang, Yang Liu, Chenguang Zhu, Michael Zeng, Cha Zhang, and Mohit Bansal. 2023. Unifying vision, text, and layout for universal document processing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 19254–19264. IEEE.

Eshaan Tanwar, Subhabrata Dutta, Manish Borthakur, and Tanmoy Chakraborty. 2023. Multilingual LLMs are better cross-lingual in-context learners with alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 6292–6307. Association for Computational Linguistics.

Suzanne Liebowitz Taylor, Deborah A Dahl, Mark Lipshutz, Carl Weir, Lewis M Norton, Roslyn Nilson, and Marcia C Linebarger. 1994. Integrated text and image understanding for document understanding. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. 2022. Hierarchical multimodal transformers for multi-page DocVQA. *CoRR*, abs/2212.05935.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Mohit Tuteja and Daniel González Juclà. 2023. Long text classification using transformers with paragraph selection strategies. In *Proceedings of the Natural Legal Language Processing Workshop 2023*, pages 17–24.

Daniel Varab and Natalie Schluter. 2021. MassiveSumm: a very large-scale, very multilingual, news summarisation dataset. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 10150–10161. Association for Computational Linguistics.

Matheus Palhares Viana and Dário Augusto Borges Oliveira. 2017. Fast CNN-based document layout analysis. In *2017 IEEE International Conference on Computer Vision Workshops, ICCV Workshops 2017, Venice, Italy, October 22-29, 2017*, pages 1173–1180. IEEE Computer Society.

Xuan-Son Vu, Quang-Anh Bui, Nhu-Van Nguyen, Thi Tuyet Hai Nguyen, and Thanh Vu. 2021. MC-OCR challenge: Mobile-captured image document recognition for Vietnamese receipts. In *RIVF International Conference on Computing and Communication Technologies, RIVF 2021, Hanoi, Vietnam, August 19-21, 2021*, pages 1–6. IEEE.

Chenglong Wang, Kedar Tatwawadi, Marc Brockschmidt, Po-Sen Huang, Yi Mao, Oleksandr Polozov, and Rishabh Singh. 2018. Robust text-to-SQL generation with execution-guided decoding. *arXiv preprint arXiv:1807.03100*.

Dongsheng Wang, Natraj Raman, Mathieu Sibue, Zhiqiang Ma, Petr Babkin, Simerjot Kaur, Yulong Pei, Armineh Nourbakhsh, and Xiaomo Liu. 2024a. DocLLM: A layout-aware generative language model for multimodal document understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 8529–8548. Association for Computational Linguistics.

Jiapeng Wang, Lianwen Jin, and Kai Ding. 2022a. LiLT: A simple yet effective language-independent layout transformer for structured document understanding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 7747–7757. Association for Computational Linguistics.

Jiapeng Wang, Chongyu Liu, Lianwen Jin, Guozhi Tang, Jiaxin Zhang, Shuaitao Zhang, Qianying Wang, Yaqiang Wu, and Mingxiang Cai. 2021a. Towards robust visual information extraction in real world: New dataset and novel solution. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 2738–2745. AAAI Press.

Jiawei Wang, Kai Hu, and Qiang Huo. 2024b. DLAFormer: An end-to-end transformer for document layout analysis. In *Document Analysis and Recognition - ICDAR 2024 - 18th International Conference, Athens, Greece, August 30 - September 4, 2024, Proceedings, Part IV*, volume 14807 of *Lecture Notes in Computer Science*, pages 40–57. Springer.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024c. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *CoRR*, abs/2409.12191.

Runchuan Wang, Zhao Zhang, Fuzhen Zhuang, Dehong Gao, Yi Wei, and Qing He. 2021b. Adversarial domain adaptation for cross-lingual information retrieval with multilingual BERT. In *CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021*, pages 3498–3502. ACM.

Wenjin Wang, Yunhao Li, Yixin Ou, and Yin Zhang. 2023a. Layout and task aware instruction prompt for zero-shot document image question answering. *CoRR*, abs/2306.00526.

Yau-Shian Wang, Ashley Wu, and Graham Neubig. 2022b. English contrastive learning can learn universal cross-lingual sentence embeddings. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 9122–9133. Association for Computational Linguistics.

Yi Wang and Xiao-Jing Wang. 2005. A new approach to feature selection in text classification. In *2005 International conference on machine learning and cybernetics*, volume 6, pages 3814–3819. IEEE.

Zilong Wang, Yiheng Xu, Lei Cui, Jingbo Shang, and Furu Wei. 2021c. LayoutReader: Pre-training of text and layout for reading order detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 4735–4744. Association for Computational Linguistics.

Zilong Wang, Yichao Zhou, Wei Wei, Chen-Yu Lee, and Sandeep Tata. 2023b. VRDU: A benchmark for visually-rich document understanding. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2023, Long Beach, CA, USA, August 6-10, 2023*, pages 5184–5193. ACM.

Zixiang Wang, Jian Yang, Tongliang Li, Linzheng Chai, Jiaheng Liu, Ying Mo, Jiaqi Bai, and Zhoujun Li. 2023c. Multilingual entity and relation extraction from unified to language-specific training. In *Proceedings of the International Conference on Electronics, Computers and Communication Technology, CECCT 2023, Guilin, China, November 17-19, 2023*, pages 98–105. ACM.

Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, et al. 2024. ChatIE: Zero-shot information extraction via chatting with ChatGPT. *arXiv preprint arXiv:2302.10205*.

Chenxi Whitehouse, Fantine Huot, Jasmijn Bastings, Mostafa Dehghani, Chu-Cheng Lin, and Mirella Lapata. 2024. Low-rank adaptation for multilingual summarization: An empirical study. In *Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 1202–1228. Association for Computational Linguistics.

Sam Wiseman, Stuart M Shieber, and Alexander M Rush. 2017. Challenges in data-to-document generation. *arXiv preprint arXiv:1707.08052*.

Wen Xiao, Iz Beltagy, Giuseppe Carenini, and Arman Cohan. 2022. PRIMERA: pyramid-based masked sentence pre-training for multi-document summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 5245–5263. Association for Computational Linguistics.

Yong Xie, Karan Aggarwal, and Aitzaz Ahmad. 2024. Efficient continual pre-training for building domain specific large language models. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 10184–10201. Association for Computational Linguistics.

Hangdi Xing, Changxu Cheng, Feiyu Gao, Zirui Shao, Zhi Yu, Jiajun Bu, Qi Zheng, and Cong Yao. 2024. DocHieNet: A large and diverse dataset for document hierarchy parsing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 1129–1142. Association for Computational Linguistics.

Peng Xu, Wei Ping, Xianchao Wu, Lawrence McAfee, Chen Zhu, Zihan Liu, Sandeep Subramanian, Evelina Bakhturina, Mohammad Shoeybi, and Bryan Catanzaro. 2024a. Retrieval meets long context large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Ruochen Xu, Chenguang Zhu, Yu Shi, Michael Zeng, and Xuedong Huang. 2020a. Mixed-lingual pre-training for cross-lingual summarization. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, AACL/IJCNLP 2020, Suzhou, China, December 4-7, 2020*, pages 536–541. Association for Computational Linguistics.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei A. F. Florêncio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. 2021a. LayoutLMv2: Multi-modal pre-training for visually-rich document understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 2579–2591. Association for Computational Linguistics.

Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. 2020b. LayoutLM: Pre-training of text and layout for document image understanding. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, pages 1192–1200. ACM.

Yiheng Xu, Tengchao Lv, Lei Cui, Guoxin Wang, Yijuan Lu, Dinei Florêncio, Cha Zhang, and Furu Wei. 2021b. Layoutxlm: Multimodal pre-training for multilingual visually-rich document understanding. *CoRR*, abs/2104.08836.

Yiheng Xu, Tengchao Lv, Lei Cui, Guoxin Wang, Yijuan Lu, Dinei A. F. Florêncio, Cha Zhang, and Furu Wei. 2022. XFUND: A benchmark dataset for multilingual visually rich form understanding. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3214–3224. Association for Computational Linguistics.

Yuemei Xu, Ling Hu, Jiayi Zhao, Zihan Qiu, Yuqi Ye, and Hanwen Gu. 2024b. A survey on multilingual large language models: Corpora, alignment, and bias. *CoRR*, abs/2404.00929.

Kota Yamaguchi. 2021. Canvasvae: Learning to generate vector graphic documents. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 5461–5469. IEEE.

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. 2024. Corrective retrieval augmented generation. *CoRR*, abs/2401.15884.

Eugene Yang, Suraj Nair, Dawn J. Lawrie, James Mayfield, and Douglas W. Oard. 2022. Parameter-efficient zero-shot transfer for cross-language dense retrieval with adapters. *CoRR*, abs/2212.10448.

Hongyang Yang, Boyu Zhang, Neng Wang, Cheng Guo, Xiaoli Zhang, Likun Lin, Junlin Wang, Tianyu Zhou, Mao Guan, Runjia Zhang, and Christina Dan Wang. 2024. Finrobot: An open-source AI agent platform for financial applications using large language models. *CoRR*, abs/2405.14767.

Huichen Yang and William H. Hsu. 2022. Transformer-based approach for document layout understanding. In *2022 IEEE International Conference on Image Processing, ICIP 2022, Bordeaux, France, 16-19 October 2022*, pages 4043–4047. IEEE.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2369–2380. Association for Computational Linguistics.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. DocRED: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 764–777. Association for Computational Linguistics.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. MiniCPM-V: A GPT-4V level MLLM on your phone. *CoRR*, abs/2408.01800.

Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Yuhao Dan, Chenlin Zhao, Guohai Xu, Chenliang Li, Junfeng Tian, et al. 2023. mPLUG-DocOwl: Modularized multimodal large language model for document understanding. *arXiv preprint arXiv:2307.02499*.

Yangfan Ye, Xiachong Feng, Xiaocheng Feng, Weitao Ma, Libo Qin, Dongliang Xu, Qing Yang, Hongtao Liu, and Bing Qin. 2024. GlobeSumm: A challenging benchmark towards unifying multi-lingual, cross-lingual and multi-document news summarization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 10803–10821. Association for Computational Linguistics.

Xiaohan Yi, Liangcai Gao, Yuan Liao, Xiaode Zhang, Runtao Liu, and Zhuoren Jiang. 2017. CNN based page object detection in document images. In *14th IAPR International Conference on Document Analysis and Recognition, ICDAR 2017, Kyoto, Japan, November 9-15, 2017*, pages 230–235. IEEE.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, et al. 2024a. VisRAG: Vision-based retrieval-augmented generation on multi-modality documents. *CoRR*, abs/2410.10594.

Wenhao Yu, Hongming Zhang, Xiaoman Pan, Peixin Cao, Kaixin Ma, Jian Li, Hongwei Wang, and Dong Yu. 2024b. Chain-of-note: Enhancing robustness in retrieval-augmented language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 14672–14685. Association for Computational Linguistics.

Wenwen Yu, Ning Lu, Xianbiao Qi, Ping Gong, and Rong Xiao. 2020. PICK: processing key information extraction from documents using improved graph learning-convolutional networks. In *25th International Conference on Pattern Recognition, ICPR 2020, Virtual Event / Milan, Italy, January 10-15, 2021*, pages 4363–4370. IEEE.

Yuechen Yu, Yulin Li, Chengquan Zhang, Xiaoqiang Zhang, Zengyuan Guo, Xiameng Qin, Kun Yao, Junyu Han, Errui Ding, and Jingdong Wang. 2023. StrucTexTv2: Masked visual-textual prediction for document image pre-training. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Klim Zaporojets, Johannes Deleu, Chris Develder, and Thomas Demeester. 2021. DWIE: an entity-centric dataset for multi-task document-level information extraction. *Inf. Process. Manag.*, 58(4):102563.

Chong Zhang, Yi Tu, Yixi Zhao, Chenshu Yuan, Huan Chen, Yue Zhang, Mingxu Chai, Ya Guo, Huijia Zhu, Qi Zhang, and Tao Gui. 2024a. Modeling layout reading order as ordering relations for visually-rich document understanding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 9658–9678. Association for Computational Linguistics.

Fangfang Zhang, Jin-ge Yao, and Rui Yan. 2018. On the abstractiveness of neural document summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 785–790. Association for Computational Linguistics.

Kechi Zhang, Jia Li, Ge Li, Xianjie Shi, and Zhi Jin. 2024b. CodeAgent: Enhancing code generation with tool-integrated agent systems for real-world repo-level coding challenges. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 13643–13658. Association for Computational Linguistics.

Peng Zhang, Yunlu Xu, Zhanzhan Cheng, Shiliang Pu, Jing Lu, Liang Qiao, Yi Niu, and Fei Wu. 2020. TRIE: end-to-end text reading and information extraction for document understanding. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1413–1422.

Ruochen Zhang and Carsten Eickhoff. 2024. CroCoSum: A benchmark dataset for cross-lingual code-switched summarization. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 4113–4126. ELRA and ICCL.

Shaolei Zhang, Kehao Zhang, Qingkai Fang, Shoutao Guo, Yan Zhou, Xiaodong Liu, and Yang Feng. 2024c. BayLing 2: A multilingual large language model with efficient language alignment. *CoRR*, abs/2411.16300.

Tianjun Zhang, Shishir G. Patil, Naman Jain, Sheng Shen, Matei Zaharia, Ion Stoica, and Joseph E. Gonzalez. 2024d. RAFT: adapting language model to domain specific RAG. *CoRR*, abs/2403.10131.

Wenxuan Zhang, Hou Pong Chan, Yiran Zhao, Mahani Aljunied, Jianyu Wang, Chaoqun Liu, Yue Deng, Zhiqiang Hu, Weiwen Xu, Yew Ken Chia, Xin Li, and Lidong Bing. 2024e. SeaLLMs 3: Open foundation and chat multilingual large language models for southeast asian languages. *CoRR*, abs/2407.19672.

Xiang Zhang, Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. 2023. Don't trust ChatGPT when your question is not in English: A study of multilingual abilities and types of LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 7915–7927. Association for Computational Linguistics.

Xingxing Zhang, Furu Wei, and Ming Zhou. 2019. HIBERT: document level pre-training of hierarchical bidirectional transformers for document summarization. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5059–5069. Association for Computational Linguistics.

Yuanchi Zhang, Yile Wang, Zijun Liu, Shuo Wang, Xiaolong Wang, Peng Li, Maosong Sun, and Yang Liu. 2024f. Enhancing multilingual capabilities of large language models through self-distillation from resource-rich languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 11189–11204. Association for Computational Linguistics.

Zhihan Zhang, Dong-Ho Lee, Yuwei Fang, Wenhao Yu, Mengzhao Jia, Meng Jiang, and Francesco Barbieri. 2024g. PLUG: leveraging pivot language in cross-lingual instruction tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 7025–7046. Association for Computational Linguistics.

Ruochen Zhao, Hailin Chen, Weishi Wang, Fangkai Jiao, Do Xuan Long, Chengwei Qin, Bosheng Ding, Xiaobao Guo, Minzhi Li, Xingxuan Li, and Shafiq Joty. 2023. Retrieving multimodal information for augmented generation: A survey. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 4736–4756. Association for Computational Linguistics.

Wei Zhao, Steffen Eger, Johannes Bjerva, and Isabelle Augenstein. 2021. Inducing language-agnostic multilingual representations. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics, *SEM 2021, Online, August 5-6, 2021*, pages 229–240. Association for Computational Linguistics.

Xiaohui Zhao, Zhuo Wu, and Xiaoguang Wang. 2019. CUTIE: learning to understand documents with convolutional universal text information extractor. *CoRR*, abs/1903.12363.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024a. How do large language models handle multilingualism? *CoRR*, abs/2402.18815.

Yiran Zhao, Wenxuan Zhang, Huiming Wang, Kenji Kawaguchi, and Lidong Bing. 2024b. Adamergex: Cross-lingual transfer with large language models via adaptive adapter merging. *CoRR*, abs/2402.18913.

Zhiyuan Zhao, Hengrui Kang, Bin Wang, and Conghui He. 2024c. DocLayout-YOLO: Enhancing document layout analysis through diverse synthetic data and global-to-local adaptive perception. *CoRR*, abs/2410.12628.

Xinru Zheng, Xiaotian Qiao, Ying Cao, and Rynson W. H. Lau. 2019. Content-aware generative modeling of graphic design layouts. *ACM Trans. Graph.*, 38(4):133:1–133:15.

Victor Zhong, Caiming Xiong, and Richard Socher. 2017. Seq2SQL: Generating structured queries from natural language using reinforcement learning. *arXiv preprint arXiv:1709.00103*.

Xu Zhong, Jianbin Tang, and Antonio Jimeno-Yepes. 2019. PubLayNet: Largest dataset ever for document layout analysis. In *2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, September 20-25, 2019*, pages 1015–1022. IEEE.

Fengbin Zhu, Wenqiang Lei, Chao Wang, Jianming Zheng, Soujanya Poria, and Tat seng Chua. 2021. Retrieving and reading: A comprehensive survey on open-domain question answering. *ArXiv*, abs/2101.00774.

Fengbin Zhu, Chao Wang, Fuli Feng, Zifeng Ren, Moxin Li, and Tat-Seng Chua. 2024a. Doc2SoarGraph: Discrete reasoning over visually-rich table-text documents via semantic-oriented hierarchical graphs. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5119–5131, Torino, Italia. ELRA and ICCL.

Junnan Zhu, Qian Wang, Yining Wang, Yu Zhou, Jiajun Zhang, Shaonan Wang, and Chengqing Zong. 2019. NCLS: neural cross-lingual summarization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 3052–3062. Association for Computational Linguistics.

Wenhao Zhu, Yunzhe Lv, Qingxiu Dong, Fei Yuan, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Extrapolating large language models to non-English by aligning languages. *CoRR*, abs/2308.04948.

Zichen Zhu, Yang Xu, Lu Chen, Jingkai Yang, Yichuan Ma, Yiming Sun, Hailin Wen, Jiaqi Liu, Jinyu Cai, Yingzi Ma, Situo Zhang, Zihan Zhao, Liangtai Sun, and Kai Yu. 2024b. MULTI: multimodal understanding leaderboard with text and images. *CoRR*, abs/2402.03173.

Ran Zmigrod, Pranav Shetty, Mathieu Sibue, Zhiqiang Ma, Armineh Nourbakhsh, Xiaomo Liu, and Manuela Veloso. 2024. "What is the value of templates?" Rethinking document information extraction datasets for LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13162–13185, Miami, Florida, USA. Association for Computational Linguistics.

A Evolution: Early Approaches

In this section, we review early multimodal and multilingual approaches for document-related understanding and generation tasks.

A.1 Understanding Tasks

A.1.1 Multimodal

Before the rise of LLMs, multimodal document understanding relied heavily on rule-based systems, template matching, and traditional machine learning. Key early models integrate text and layout information using feature engineering techniques.

Rule-based systems rely on manually defined sets of rules and logic to interpret and process data. Niyogi and Srihari (1986) introduces a rule-based system that analyzes digitized document images. The system identifies various printed regions within the document and classifies them into logical "blocks" of information. It employs a goal-directed, top-down approach, utilizing a three-level rule hierarchy to guide its control strategy. Klink and Kieninger (2001) presents a rule-based system that combines layout and textual features to understand document structure. The system uses fuzzy logic to merge these features, allowing for adaptable rule bases that can be tailored to new domains. It processes documents in two stages: document analysis and document understanding. Markovic and Gostojic (2020) proposes a knowledge-based method for assembling legal documents. It utilizes a machine-readable representation of legal professionals' knowledge, comprising a rule base for formal legal norms and document templates for tacit knowledge. The system collects input data through an interactive interview, performs legal reasoning over this data, and generates the output document. It also creates an argument graph to explain the reasoning process, providing users with an interpretation of how input data and the rule base influence the output document.

Template Matching is a technique where predefined patterns or "templates" are compared to input data (such as text or images) to identify similarities. Dhakal et al. (2019) introduces a novel one-shot template-matching algorithm designed to automatically capture data from business documents, aiming to minimize manual data entry. The algorithm utilizes a set of engineered visual and textual features, enabling it to be invariant to changes in position and value. Nguyen Vu et al. (2024) presents a component-based template matching approach that leverages document layout analysis. This method involves sub-graph mining to extract significant sub-graphs from a training set, identifying recurrent patterns. A deep graph neural network is then employed to align these sub-graphs with segments in the query document, enhancing the robustness of the template matching algorithm.

Traditional Machine Learning involves algorithms that learn patterns from data without being explicitly programmed with rules. In the visual domain, Viana and Oliveira (2017) and Yi et al. (2017) first introduce R-CNN (Girshick et al., 2014) into DLA task, followed by studies (Biswas et al., 2022; Saha et al., 2019; Yang and Hsu, 2022; Wang et al., 2024b) adopting more advanced detection and segmentation models. The progress also fuels research interest in applications for diverse document types and specific document components such as element segmentation for newspaper (Almutairi and Alfarhood, 2019) and historical documents (Oliveira et al., 2018), keyword detection in Chinese documents (Lin et al., 2021) and graphical object detection (Saha et al., 2019) for mathematical equations and figures. To handle lengthy text, LSTM and transformer module are widely adopted (Wang et al., 2021c; Rausch et al., 2023; Wang et al., 2024b). More recent works design specific module to interact with modalities uniquely present in documents, such as layout and chart. Chargrid (Katti et al., 2018) is one of the first models to process documents as 2D grids, treating document content as both text and spatial layout. TRIE (Zhang et al., 2020) combined text recognition with information extraction, leveraging end-to-end methods but still using handcrafted features for layout understanding. Yu et al. (2020) and Liu et al. (2019a) treat documents as a graph and embed width and height of text segment as well as vertical and horizontal distance between two segments as graph embedding. With the advancement of the transformer-based model, coordinate encoding gradually becomes the default paradigm. LayoutLM (Xu et al., 2020b) pre-trains BERT models with coordinate information and image embeddings integrated with text and positional embeddings. Follow-up works either incorporate additional pre-training tasks (Xu et al., 2021a; Hong et al., 2022; Huang et al., 2022) and reading order (Gu et al., 2022) from layout information or inject 2D token position into transformer's attention bias (Powalski et al., 2021; Garncarek et al., 2021; Tang et al., 2023). DocMamba (Hu et al., 2024c) explores Mamba (Gu and Dao, 2023) in DAI tasks, making use of layout segment and token coordinates to derive 1D text tokens and 2D position embeddings.

A.1.2 Multilingual

Language Model Pre-training leverages advancements in transformer-based unsupervised pre-training on large-scale multilingual datasets, employing a training objective specifically designed to optimize performance on multilingual tasks. Chi et al. (2021a) pre-trained a cross-lingual model InfoXLM by maximizing the mutual information at multi-granularity level, including monolingual token-sequence, cross-lingual token-sequence and cross-lingual sequence-level. LayoutXLM (Xu et al., 2021b) combines the multimodal and multilingual pre-training techniques from LayoutLMv2 (Xu et al., 2021a) and InfoXLM respectively. During the pre-training stage, aside from text-image alignment and matching, a novel multilingual masked visual-language modeling objective is proposed. Evaluation on language-specific finetuning, zero-shot transfer learning and multitask fine-tuning showed significant improvement over XLM-RoBERTa and InfoXLM. In the cross-lingual information retrieval (CLIR) task, Wang et al. (2021b) fine-tunes mBERT (Devlin et al., 2019) with deep relevant matching objective on a home-made CLIR training data derived from parallel corpora, showing promising results on the retrieval of Lithuanian documents against short English queries.

Language-specific Training The imbalance of data distribution in multilingual training often leads to poorer performance on Low-resource Languages (LRLs) than High-resource Languages(HRLs) (Li et al., 2024f). While the models are trained on a variety of languages, the performance in downstream tasks declines, a phenomenon known as the curse of multilinguality (Chang et al., 2024). ColBERT-XM (Louis et al., 2024) applies language-specific adapters in pre-training. Adapters and embedding layers were frozen during monolingual fine-tuning to force learning in the shared weights from high-resourced languages. Low-resource languages can benefit from data-efficient training via updating their adapters only. Similar work (Yang et al., 2022) on CLIR shows that models trained with monolingual data are more effective than fine-tuning the entire model.

Language-independent Approach assumes features other than text could remain invariant across documents in different languages. LiLT (Wang et al., 2022a) separates text and layout features during the pre-training stage, enhancing layout structure learning using monolingual data. Multilingual fine-tuning further boosts cross-lingual capabilities while requiring less supervised data. Inspired by LiLT, Shen et al. (2024) further decouples visual features during pre-training. Since image contains language information, they adopt a text edit diffusion model to replace the original texts with fictional words.

Translation-aided Approach Liu et al. (2021) introduces a data augmentation framework via translation for low-resource named entity recognition (NER) task. The sentences are input into the translation system in two modified forms of the original source sentence. In the first form, entity words are replaced with corresponding entity tags and indexes. In the second form, brackets are used to include word span information alongside the entity words. After translation, the entity tags in the output are replaced with the translations extracted from the bracketed spans.

A.2 Generation Tasks

A.2.1 Multimodal

While effective for understanding document structure, **statistical methods** often fall short in generating new content, as they primarily rely on existing textual patterns. Instead, early works reply to the following approaches.

Heuristic-based Approaches rely on manually designed rules or strategies to generate content. To generate summarization, these approaches analyze specific text features (Carbonell and Goldstein, 1998), such as keyword frequency (Luhn, 1958) (using metrics like TF or TF-IDF to identify key sentences), sentence length, positional importance (prioritizing content at the beginning or end of a document), and title matching (selecting sentences with title-related keywords) (Edmundson, 1969). To generate answers, heuristic systems answer the questions from predefined patterns within a closed domain (Gupta and Gupta, 2012). Similarly, to generate document classification tags, feature engineering techniques like tokenization, stop-word removal, and stemming (Wang and Wang, 2005) are used to identify linguistic patterns. While straightforward, these methods often lack the flexibility of more advanced approaches.

Optimization-based Approaches aim to mathematically identify the "best" unseen content under specific constraints. To generate summarization, these methods optimize objectives such as maximizing information coverage (Takamura and Okumura, 2009) or minimizing redundancy (Hirao et al., 2013) to achieve an ideal compression ratio. Some notable techniques include linear programming (McDonald, 2007) for selecting sentence subsets that satisfy predefined criteria and sub-modular optimization (Lin and Bilmes, 2011) to balance diversity and coverage. In question answering, optimization frameworks enhance the precision and recall of retrieved answers, while in document classification, they are used to refine feature selection and representation, enabling the effective handling of lengthy and complex documents.

Traditional Machine Learning has advanced document generation through deep learning and pre-trained models. To generate summarization, supervised learning trains neural networks on document-summary pairs (Nallapati et al., 2017; Narayan et al., 2018; See et al., 2017), using word embeddings like Word2Vec (Mikolov, 2013) and GloVe (Pennington et al., 2014). To generate answers against document-related questions, more modalities are taken into consideration, such as images and charts. Datasets like OCR-VQA (Mishra et al., 2019), TextVQA(Singh et al., 2019), and DocVQA(Mathew et al., 2021) focus on single-image QA, while MultiModalQA (Talmor et al., 2021) and MP-DocVQA (Tito et al., 2022) emphasize reasoning across multiple pages or images. Generative models in table-based QA directly generate answers, such as Seq2Seq (Dong and Lapata, 2016; Zhong et al., 2017; Wang et al., 2018) and graph-based approaches (Mueller et al., 2019). Seq2Seq models focusing on non-database tables and free-form QA. Graph-based methods represent tables as graphs but are limited to table-only tasks. To generate document tags for classification, machine learning pipelines (Han et al., 2024) utilize tokenization, vectorization, and embedding to model high-dimensional text data. Neural networks, particularly transformers, excel at handling lengthy and multi-paragraph documents (Tuteja and Juclà, 2023).

A.2.2 Multilingual

Pipeline Method decomposes the original task into sub-tasks and involves machine translation either as the first step or the last step to avoid resource-intensive multilingual annotation. Leuski et al. (2003) first translates Hindi documents to English using a statistical machine translation system, after which important English sentences are selected to create a summary. Araújo et al. (2016) compares twenty-one methods and two language-specific models on nine language-specific dataset. The finding suggests that simply translating the input text to English and leveraging English-based methods is better than language-specific methods. Duh et al. (2011) argues that cross-lingual errors are not caused by machine translation errors, and would occur even with a perfect translation model. A better cross-lingual algorithm is needed to mitigate the errors.

Pre-trained Model Approach The emergence of per-trained models has benefited almost every area of NLP, including multilingual document generation tasks. By simply fine-tuning on large-scale cross-lingual summarization datasets, it's able to outperform many multi-task models (Liang et al., 2022). Similar work (Amin et al., 2023) also shows the effectiveness of mBERT on cross-lingual QA tasks.

Cross-lingual Transfer explores the language transfer capability of models with limited multilingual corpora. Lee and Lee (2019) explores cross-lingual transfer learning for question answering by leveraging a source language with abundant annotations to improve performance in a target language with limited data. A GAN-based approach is proposed to incorporate a language discriminator to learn language-universal feature representations, and consequentially transfer knowledge from the source language. Xu et al. (2020a) proposes a mixed-lingual pre-training approach that leverages both cross-lingual tasks, such as translation, and monolingual tasks, like masked language modeling, to enhance cross-lingual summarization. The model effectively utilizes massive monolingual data to improve language modeling and achieves significant improvements in ROUGE (Lin, 2004) scores over state-of-the-art results on the NCLS dataset (Zhu et al., 2019).

B Benchmark

The rapid advancement in DAI would not be possible without the variety of carefully curated datasets. We summarize them in Table 1 along 5 aspects: language, quantity, open-source availability, modality, and task. We hope that this information could make a valuable contribution to future multimodal and multilingual research in DAI.

C Usage of AI Assistant

We mainly use AI assistant or tools such as ChatGPT and Grammarly to check the grammar errors.

					Modality			Task						
Dataset	Language	#Languages	#Documents	Open-Source	Text	Visual	Layout	KIE	DLA	DSA	DC	DS	DCG	QA
SROIE (Huang et al., 2019)	EN	1	973	✓	✓	✓	✓	✓	✓					
FUNSD (Jaume et al., 2019)	EN	1	199	✓	✓	✓	✓	✓						
CORD (Park et al., 2019)	EN	1	1,000	✓	✓	✓	✓	✓	✓					
CUTIE (Zhao et al., 2019)	ES	1	4,484	✘	✓	✓	✓	✓						
Kleister NDA (Stanislawek et al., 2021)	EN	1	540	✓	✓	✓	✓	✓						
Kleister Charity (Stanislawek et al., 2021)	EN	1	2,788	✓	✓	✓	✓	✓						
DWIE (Zaporojets et al., 2021)	EN	1	802	✓	✓	✓	✓	✓	✓					
WildReceipt (Sun et al., 2021)	EN	1	1,768	✓	✓	✓	✓	✓	✓					
MC-OCR (Vu et al., 2021)	Multi	2	2,436	✓	✓	✓	✓	✓						
EPHOIE (Wang et al., 2021a)	CN	1	1,494	✓	✓	✓	✓	✓						
XFUND (Xu et al., 2022)	Multi	7	1,393	✓	✓	✓	✓	✓						
VRDU Registration (Wang et al., 2023b)	EN	1	1,915	✓	✓	✓	✓	✓						
VRDU Ad-buy (Wang et al., 2023b)	EN	1	641	✓	✓	✓	✓	✓						
DocILE (Simsa et al., 2023)	EN	1	6,680	✓	✓	✓	✓	✓						
POIE (Kuang et al., 2023)	EN	1	3,000	✓	✓	✓	✓	✓						
AMuRD (Abdallah et al., 2023)	Multi	2	47,720	✓	✓			✓			✓			
UIT-MLReceipts (Nguyen, 2024)	Multi	2	2,147	✘	✓	✓	✓	✓						
CORU (Abdallah et al., 2024a)	Multi	2	20,000	✓	✓	✓	✓	✓	✓					
DocRED (Yao et al., 2019)	EN	1	5,053	✓	✓	✓	✓		✓					
PubLayNet (Zhong et al., 2019)	N/A	N/A	358,353	✓	✓	✓	✓		✓					
TableBank (Li et al., 2020a)	Multi	N/A	18,000	✓	✓	✓	✓		✓					
DocBank (Li et al., 2020b)	EN	1	500,000	✓	✓	✓	✓		✓					
ReadingBank (Wang et al., 2021c)	EN	1	500,000	✓	✓	✓	✓		✓					
arXivdocs-target (Rausch et al., 2021)	EN	1	362	✓	✓	✓	✓		✓					
DocLayNet (Pfitzmann et al., 2022)	Multi	4	80,863	✓	✓	✓	✓		✓					
HRDoc (Ma et al., 2023b)	EN	1	2,500	✓	✓	✓	✓		✓					
E-Periodica (Rausch et al., 2023)	Multi	4	542	✓	✓	✓	✓		✓					
DocHieNet (Xing et al., 2024)	Multi	2	1,673	✓	✓	✓	✓		✓					
Amazon Sentiment Polarity (Prettenhofer and Stein, 2010)	Multi	4	6,000	✓	✓					✓				
Sinhala-English-Code-Switched-Dataset (Rathnayake et al., 2022)	Multi	2	10,000	✓	✓					✓				
AfriSenti (Muhammad et al., 2023)	Multi	14	112,506	✓	✓					✓				
Tobacco-3482 (Kumar et al., 2014)	EN	1	3,482	✓		✓					✓			
RVL-CDIP (Harley et al., 2015)	EN	1	400,000	✓		✓					✓			
MLDoc (Schwenk and Li, 2018)	Multi	8	6,000	✓	✓						✓			
MultiEURLEX (Chalkidis et al., 2021)	Multi	23	65,000	✓	✓						✓			
En2ZhSum (Zhu et al., 2019)	Multi	2	370,687	✓	✓							✓		
Zh2EnSum (Zhu et al., 2019)	Multi	2	1,699,713	✓	✓							✓		
MLSUM (Scialom et al., 2020)	Multi	5	1,571,041	✓	✓							✓		
Wikilingua (Ladhak et al., 2020)	Multi	18	770,000	✓	✓							✓		
XL-Sum (Hasan et al., 2021)	Multi	44	1,005,292	✓	✓							✓		
MassiveSumm (Varab and Schluter, 2021)	Multi	92	31,940,180	✓	✓							✓		
CrossSum (Bhattacharjee et al., 2023)	Multi	45	1,678,466	✓	✓							✓		
CroCoSum (Zhang and Eickhoff, 2024)	Multi	2	42,000	✓	✓							✓		
GlobeSumm (Ye et al., 2024)	Multi	26	4,687	✓	✓							✓		
Crello (Yamaguchi, 2021)	N/A	N/A	23,322	✓		✓	✓						✓	
PubGenNet (Biswas et al., 2024)	N/A	N/A	346,948	✘	✓	✓	✓						✓	
TriviaQA (Joshi et al., 2017)	EN	1	662,659	✓	✓									✓
HotpotQA (Yang et al., 2018)	EN	1	112,779	✓	✓									✓
NarrativeQA (Kociský et al., 2018)	EN	1	1,572	✓	✓									✓
TextVQA (Singh et al., 2019)	EN	1	28,408	✓	✓	✓	✓							✓
OCR-VQA (Mishra et al., 2019)	EN	1	207,572	✓	✓	✓								✓
Natural Questions (Kwiatkowski et al., 2019)	EN	1	307,373	✓	✓									✓
2WikiMultiHopQA (Ho et al., 2020)	EN	1	192,606	✓	✓									✓
VisualMRC (Tanaka et al., 2021)	EN	1	10,234	✓	✓	✓	✓							✓
WebSRC (Chen et al., 2021)	EN	1	6,400	✓	✓	✓	✓							✓
MultiModalQA (Talmor et al., 2021)	EN	1	57,713	✓	✓	✓								✓
DocVQA (Mathew et al., 2021)	EN	1	12,767	✓	✓	✓	✓							✓
InforgraphicsVQA (Mathew et al., 2022)	EN	1	5,485	✓	✓	✓	✓							✓
MP-DocVQA (Tito et al., 2022)	EN	1	5,928	✓	✓	✓	✓							✓
DuReader$_{vis}$ (Qi et al., 2022)	CN	1	158,000	✓	✓	✓	✓							✓
DUDE (Landeghem et al., 2023)	EN	1	5,019	✓	✓	✓	✓							✓
SlideVQA (Tanaka et al., 2023)	EN	1	2,619	✓	✓	✓	✓							✓
xPQA (Shen et al., 2023)	Multi	12	18,000	✓	✓									✓
TAT-DQA (Zhu et al., 2024a)	EN	1	2,758	✓	✓	✓	✓							✓
MMLongbench-Doc (Ma et al., 2024b)	EN	1	135	✓	✓	✓	✓							✓
M-LongDoc (Chia et al., 2024)	EN	1	851	✓	✓	✓	✓							✓
VisDoMBench (Suri et al., 2024)	EN	1	1,277	✓	✓	✓								✓
ViDoRe (Faysse et al., 2024)	Multi	2	8,310	✓	✓	✓								✓
X-AlpacaEva (Zhang et al., 2024g)	Multi	5	4,025	✓	✓									✓
CHIC (Mahamoud et al., 2024)	Multi	5	774	✘	✓	✓	✓							✓
Belebele (Bandarkar et al., 2024)	Multi	122	109,800	✓	✓									✓
mOSCAR (Futeral et al., 2024)	Multi	163	314,827,611	✓	✓	✓								✓

Table 1: Benchmark datasets for DAI tasks. Datasets are grouped by task, followed by publication year and language.