Paper accepted to ECML-PKDD 2026

📝 NOTESBANK: Benchmarking Neural Transcription and Search for Scientific Notes Understanding

Alloy Das[1] Sanket Biswas[2] Soumitri Chattopadhyay[3] Ayush Lodh[4] Aniket Pal[5]
Priyanka Banerjee[6] Nisha Singh[4] CV Jawahar[7] Josep Llados[2] Dimosthenis Karatzas[2]

[1]Iowa State University [2]Computer Vision Centre [3]UC San Diego [4]NIT Delhi
[5]LNMIIT Jaipur [6]Habitat Lens Pvt. Ltd. [7]IIIT Hyderabad

Abstract

Understanding and reasoning over academic handwritten notes remains a challenge in document AI, particularly for mathematical equations, diagrams, and scientific notations. Existing visual question answering (VQA) benchmarks focus on printed or structured handwritten text, limiting generalization to real-world note-taking. To address this, we introduce **NOTES-BANK**, *an evaluation benchmark for* ***Neural Transcription and Search*** *in note-based question answering.* NOTES-BANK *comprises complex notes across multiple domains, requiring models to process unstructured and multimodal content. The benchmark defines two tasks: (1)* ***Evidence-Based VQA****, where models retrieve localized answers with bounding-box evidence, and (2)* ***Open-Domain VQA****, where models classify the domain before retrieving relevant documents and answers. Unlike classical Document VQA datasets relying on optical character recognition (OCR) and structured data,* NOTES-BANK *demands vision-language fusion, retrieval, and multimodal reasoning. We benchmark state-of-the-art Vision-Language Models (VLMs) and retrieval frameworks, exposing structured transcription and reasoning limitations.* NOTES-BANK *provides a rigorous evaluation with ANLS*, MRR, Recall@K, and IoU, establishing a new standard for visual document understanding and reasoning.*

1. Introduction

"What we know is a drop, what we do not know is an ocean." – Sir Isaac Newton. Scientific notes, often handwritten and informal, serve as the foundational medium through which knowledge is initially documented, refined, and communicated. These notes typically include prose, mathematical derivations, shorthand annotations, and graphical elements like diagrams, flowcharts, and equations (as illustrated in Figure 1) – forming the core of scientific discovery, engineering innovations, and academic learning. Despite their critical role, automatically interpreting and understanding handwritten scientific notes remains a formidable challenge within the field of Visual Document Understanding (VDU) [33, 57, 62, 80].

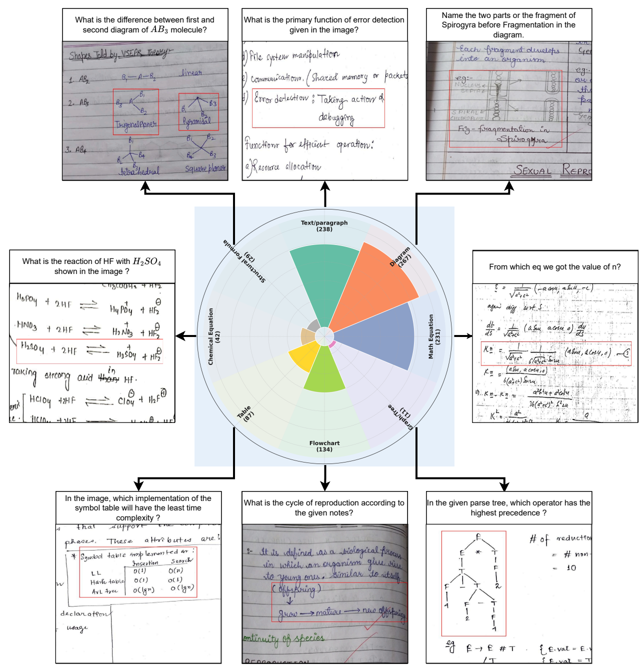

Figure 1. Samples from our introduced NOTES-BANK: *equations, flowcharts, structures, and text answers.* Center: question-type distribution. Around: annotated handwritten snippets showcasing the challenge of visual-semantic grounding.

Current VDU benchmarks do not sufficiently capture the complexities posed by handwritten scientific notes. Prior datasets either focus on neatly rendered handwriting, such as HW-SQuAD [43], or historical texts like BenthamQA, but none adequately represent modern lecture or notebook pages containing a mixture of complex graphical and textual content. Standard benchmarks for document understanding—DocVQA [42, 66], PubLayNet [79], and Do-

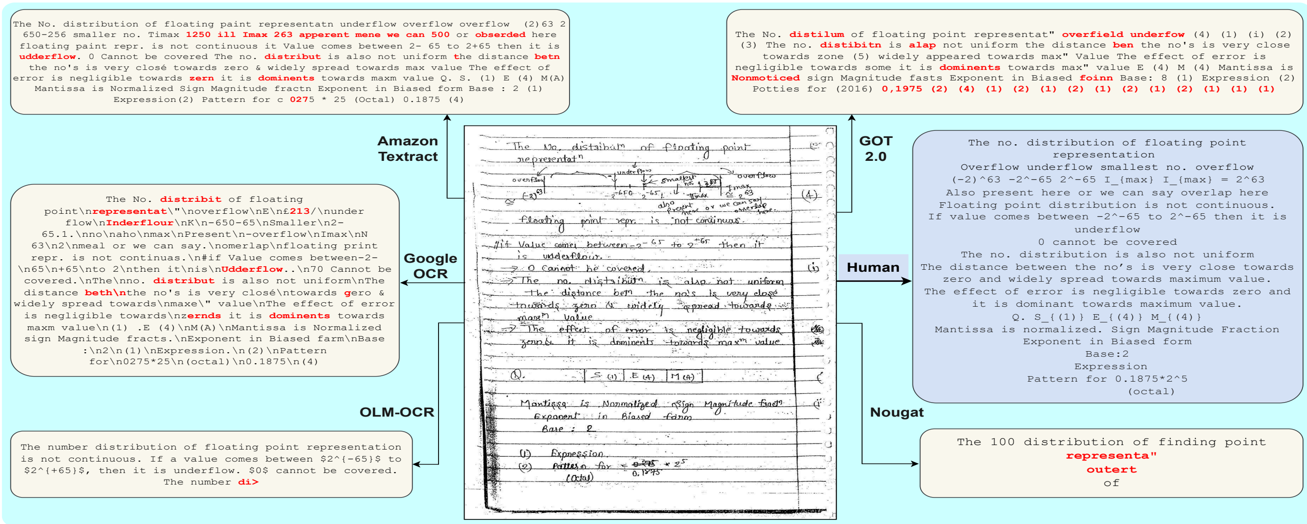

Figure 2. *Comparison of OCRs on a challenging handwritten scientific note.* Commercial engines (e.g., Textract, Google-OCR) fail on math symbols, structure, and semantics. Open-source ones (e.g., Nougat, GOT 2.0, OLM-OCR) struggle with basic text extraction. Highlights need for *multimodal reasoning beyond OCR* in handwritten understanding.

cLayNet [50]—primarily address structured layouts, where the content has well-defined segmentation and reading order. These datasets exhibit significant performance drops when tasks require interpreting scattered or non-linear layouts involving diagrams, equations, or informal handwriting. Such gaps highlight the absence of suitable benchmarks explicitly dedicated to the complexities inherent in handwritten scientific notes.

In parallel, recent advancements in vision-language models (VLMs) [29, 67] and multimodal retrieval-augmented generation (RAG) methods [12, 74] have considerably improved performance in multimodal document understanding tasks. OCR-based VDU models like LayoutLM [25, 70] and DocFormer [4] heavily rely on extracted textual content, thus losing critical visual information such as handwriting styles, layout nuances, or visual markings. Similarly, existing OCR pipelines [8, 69] typically discard visual cues during the image-to-text conversion process, reducing the models' ability to interpret complex visual structures common in notes (Figure 2). Recent analyses [18] further highlight this issue, emphasizing that current document retrieval systems primarily utilize transcribed text while neglecting essential visual context. Thus, handwritten notes fundamentally differ from structured printed documents, underscoring the need for approaches that treat document images as first-class inputs rather than relying solely on transcribed textual data.

In this paper, we introduce NoTeS-Bank, a meticulously curated benchmark designed to facilitate multimodal reasoning on unstructured scientific note documents. Our benchmark encompasses several STEM topics ranging from computer science to physics to biological sciences, with diverse visio-graphical elements such as graphs, flowcharts, mathematical equations, tables, among others (more details in upcoming sections). Figure 1 illustrates the extent of diversity these documents may have in terms of their constituent graphical elements. We also define two tasks that very well align with the data in hand. The *first* task is ***Evidence-based VQA***, which extends mere language-driven question-answering [54, 64] to a visual grounding task [14, 71], where the model is challenged to navigate through a note document, parse the unstructured handwritten text along with complex equations, multiple graphical elements, and accurately localize the answer within the document. This builds trust, explains reasoning, and validates that the model does not merely parrot patterns but end up truly "reading" the document, much like humans. The *second task* is ***Open-domain QA***, where the model is prompted with complex, open-ended questions whose answers may lie arbitrarily anywhere in a collection of multi-page notes. The model first narrows down the relevant subject pertinent to the question, searches through its collection of documents, and pieces together a consolidated answer retrieved across multiple pages.

We conduct extensive experimentation on NoTeS-Bank to evaluate (1) evidence-based reasoning and (2) retrieval-augmented open-domain question-answering abilities of a suite of recently proposed foundational multimodal language models, encompassing VLMs [30, 47, 61, 67], OCR-driven LLMs [15, 22], layout-guided LLMs [22, 34] and VLM-based OCR models [8, 51], which have previously shown promise across document reasoning tasks [12, 40, 41, 59]. Furthermore, we also establish a human performance baseline by involving qualified domain-aware individuals uninvolved in the annotation process. Employing relevant empirical metrics, our holistic evaluation re-

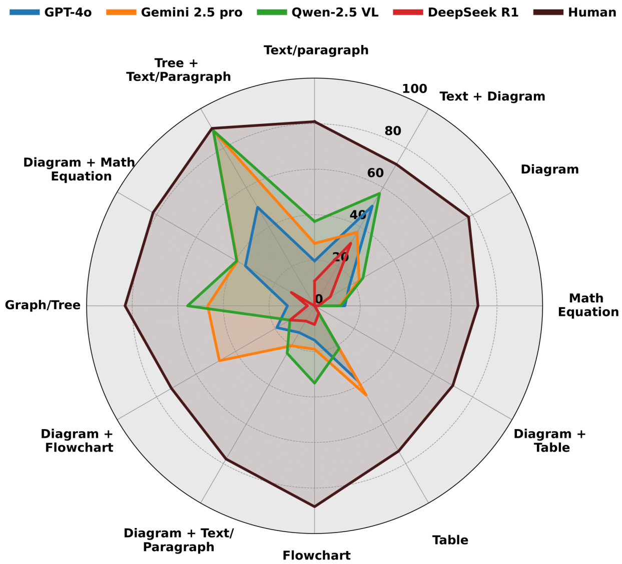

Figure 3. Scientific notes often contain *unstructured graphical elements* such as diagrams, flowcharts, graphs and equations, in addition to extractable text. Our evaluation (ANLS* Score) reveals that all existing multimodal LLMs struggle with QA when documents contain primarily untranscribable graphical elements (e.g., diagrams, equations). Performance improves when transcribable text is present.

veals that current models struggle to achieve satisfactory performance on handwritten note reasoning, showing significantly large gaps compared to the upper bound human performance (Figure 3), making NOTES-BANK a challenging benchmark that advocates further advancement of multimodal foundational model research.

Our key contributions are summarized as follows:

- We present NOTES-BANK, a novel benchmark for question answering over unstructured, scientific notes, addressing a gap in multimodal document understanding by focusing on visio-graphical content beyond printed or structured formats.
- We define two tasks—Evidence-Based VQA and Open-Domain VQA—that jointly evaluate answer grounding, domain classification, and retrieval-based multimodal reasoning in challenging handwritten scenarios.
- We benchmark a diverse set of VLMs, OCR+LLM pipelines, and retrieval-augmented approaches, and provide a comprehensive evaluation framework using popular metrics like ANLS*, IoU, Recall@K, and MRR to highlight the modality gap in current systems.

2. Related Works

Visual Question Answering. VQA provides a natural language interface for tackling diverse vision-language tasks, merging computer vision and natural language processing (NLP) techniques. This approach has been widely applied across multiple domains, including medical question answering [26, 46, 53], open-domain knowledge retrieval [37, 39, 45, 72], emotion recognition [7, 20], code-based QA [2, 35], logical reasoning [38, 75], fact verification [24, 76], and mathematical reasoning [10, 23, 40, 77].

Visual Document Understanding. The field of visual document understanding (VDU) has progressed rapidly with benchmarks that test multimodal reasoning in Document VQA models. Early datasets like DocVQA [42, 62] and InfographicsVQA [44] focused on single-page comprehension, while recent ones introduce more realistic and complex scenarios. DUDE [65, 66], a large-scale benchmark spanning multi-page, multi-domain documents, challenges models with cross-page reasoning and reveals that even state-of-the-art layout-aware Transformers [25, 70] and VLMs [1, 3, 9, 31, 32] fall significantly short of human performance [13]. Specialized tasks such as SlideVQA [60], TableVQA [28], and ScreenUI [5] further push models to understand slides, tables, and UI-like layouts. MMLongBench-Doc [41] evaluates long-form comprehension on 50-page scientific articles. MMDocBench [80], with 4,000 QA pairs across 15 diverse tasks, benchmarks VLMs like GPT-4V [1], LLaVA [36], and InternVL [11] in zero-shot, OCR-free settings. Meanwhile, datasets like VisDomRAG [59] and M3DocVQA [12] combine document images with open-domain content to test multimodal RAG systems. This expanding set of benchmarks (see Table 1) is driving progress toward robust, generalizable document understanding. Building on this, we introduce NOTES-BANK—the first benchmark focused on handwritten, unstructured academic notes—challenging models to reason without OCR or structured text cues.

3. The NOTES-BANK Benchmark Suite

NOTES-BANK is a gold-standard evaluation benchmark designed to assess multimodal question answering over complex unstructured scientific notes. Unlike existing DocVQA datasets [42–44, 62, 66], NOTES-BANK introduces two distinct tasks that challenge VLMs, multimodal LLMs and retrieval-augmented generative (RAG) architectures.

3.1. Evidence-Based VQA

The Evidence-Based VQA (EB-VQA) task in NOTES-BANK evaluates a model's ability to retrieve, comprehend, and justify answers using handwritten note-based evidence. Unlike existing DocVQA challenges that rely solely on extracted OCR text, this task requires models to reason over visual semantics, structural elements, and handwritten symbols while ensuring explicit grounding of responses.

Task Formulation: Given an input visual note (image) I (which could span 1-3 pages) containing unstructured text, symbols, equations and diagrams, and a natural language

Table 1. Comparison of benchmarks on content type, document setting, domain, and task type.

Benchmark	Content Type	Multi Document	Domain	Tasks
LongBench [6]	Text	✓	Wikipedia	Long-form QA, Retrieval
MPDocVQA [63]	Text, Tables, Charts	✗	Multi	Document Visual QA
∞-Bench [78]	Text	✗	Multi	List QA, Reasoning
DUDE [66]	Text, Tables, Charts, Figures	✗	Multi	Document Visual QA, Multi-hop QA, Unanswerable, List QA
MMLongBench-Doc [41]	Text, Tables, Charts, Slides	✗	Multi	Document QA, List QA
M3DocVQA [12]	Text, Tables, Charts	✓	Wikipedia	Open-domain Document VQA
VisDoMBench [59]	Text, Tables, Charts, Slides	✓	Multi	Evidence-based Visual Grounding with Bbox, Open-domain QA
NOTES-BANK (Ours)	**Graphical Diagram, Math Equation, Chemical Equation, Structural formula, Text/paragraph, Graph/Tree, Flowchart, Table**	✓	**Multi (Scientific), Multi-Task**	**Evidence-based Visual Grounding with Bbox and Semantic Labeling, Open-domain VQA, Multi-hop QA, Unanswerable QA, Reasoning**

question Q, the model must: (a) ***Retrieve Relevant Evidence***: Identify key portions of the visual note I that contribute to answering Q *by means of a bounding box or multiple bounding boxes*. (b) ***Generate an Answer***: Synthesize a natural language response A based on the retrieved evidence E. (c) ***Provide Justification***: Highlight the supporting evidence E in I that links to the final answer, including the corresponding visual elements (e.g., mathematical equations, chemical formulas, diagrams).
Formally, the model is defined as:

$$A, E = f_{\text{EB-QA}}(I, Q)$$

where the evidence set E consists of:

$$E = \{(B_i, L_i, G_i)\}_{i=1}^{p}$$

where B_i represents the bounding box of the relevant evidence region, L_i denotes the local category of the evidence (e.g., equation, table, diagram), G_i specifies the global category related to the document's conceptual domain (e.g., group theory, rotational mechanics).

Evaluation: To evaluate model performance, we assess answer accuracy and evidence selection quality. Answer accuracy is measured using Average Normalized Levenshtein Similarity (ANLS*) [49], while evidence selection is evaluated through Intersection-over-Union (IoU), which quantifies alignment between predicted evidence E and the ground truth E^*.

$$IoU(E, E^*) = \frac{|E \cap E^*|}{|E \cup E^*|}$$

Additionally, we measure the correctness of local and global element categorization inside, ensuring that the models retrieve not only the appropriate text regions but also the relevant semantic concepts necessary for reasoning.

Table 2. Key statistics of NOTES-BANK benchmark.

Statistic	Number
Total questions	2,000
- Questions newly annotated	2,000 (100.0%)
- For task 1	1,000
- For task 2	1,000
Unique number of Documents	649
With Distractor	2,732
Unique number of questions	1,982
Unique number of answers	1,583
- Plain text answers	1,244
- LaTeX/formula answers	758
Source	diverse online Notes repository
Maximum question length	48
Maximum answer length	32
Average question length	12.4
Average answer length	3.9

3.2. Open-Domain Question Answering

The Open-Domain QA (OD-QA) task in NOTES-BANK tests a model's ability to retrieve, reason, and answer across a large set of handwritten notes. Unlike single-document QA, it requires domain classification, document retrieval, and answer generation.

Formally, given a document collection D and a natural language question Q, the model must predict the subject category C, retrieve the most relevant document I, and generate the final answer A:

$$C = f_{\text{domain}}(Q), \quad I = f_{\text{retrieve}}(D, Q, C), \quad A = f_{\text{answer}}(I, Q)$$

where C represents the predicted subject category (e.g., physics, mathematics). I is the retrieved handwritten document. A is the generated answer.

Unlike traditional retrieval-based QA, NoTeS-Bank requires models to handle noisy, unstructured, and multimodal content, including mathematical expressions, diagrams, and scientific notations. The retrieval component is evaluated using R@1, MRR, R@5 to assess ranking quality, while answer accuracy is measured through ANLS*.

3.3. Dataset Collection and Annotation

Data Collection. The dataset was collected by scraping handwritten STEM PDFs from public repositories, with *explicit permission* from each owner, filtering out low-quality, multilingual, or low-resolution scans.

Note: The original URLs are withheld for double-blind compliance (potentially revealing nationalities); they will appear in the supplementary upon acceptance.

3.3.1. Annotation Process

The annotation process is conducted in three distinct stages: the `first stage` involves the collection of relevant documents sourced from online repositories with the owners' consent, focusing on notes for entrance, board, and government exams while ensuring the exclusion of low-quality materials; the `second stage` encompasses formulating questions and answers, delineating bounding boxes to highlight pertinent evidence, and documenting associated metadata, with distinguished undergraduates from top competitive exams participating in rigorous sessions to ensure question quality and a verification process for crafting intricate questions; and finally, the `third stage` entails a meticulous verification process with multiple iterations to provide comprehensive quality checks.

To maintain the challenge and novelty of the NoTeS-Bank benchmark, we used **`Adversarial Filtering`**. This process involved generating numerous questions, which were assessed using the GPT-4o model [47] to set a performance baseline. Previous studies [21], [68] have utilized a comparative procedure to enhance the robustness of their proposed benchmarks. Questions the model answered correctly were removed, while those it couldn't answer were retained, creating a collection highlighting AI limitations and raising the difficulty standard.

Task 1: Evidence-Based QA Annotation. For the Evidence-Based QA task, annotators first created question-answer pairs by formulating queries that required reasoning over multimodal content. Each QA pair was recorded along with metadata, including if the document was single-page or multi-page, the page number where the answer appeared, and the subject name. To further categorize the nature of the answer, the ROI from which the answer was derived was labeled as a *local category*, such as text, equation, diagram, or chemical formula. Each QA pair was also assigned a *global category* corresponding to its conceptual domain, such as group theory or rotational mechanics.

Once the QA pairs were created, annotators manually labeled the corresponding answer regions by drawing bounding boxes around the relevant content using an annotation tool. These bounding boxes, along with the associated QA pairs and document images, were compiled into a structured JSON format for experimentation and evaluation.

Task 2: Open-Domain QA Annotation. A separate annotation process was conducted for the Open-Domain QA task to ensure that retrieval-based reasoning was accurately represented. A new set of QA pairs was created to require retrieval across multiple documents rather than direct extraction from a single page. In this task, the domain classification of each QA pair was explicitly recorded to assist in retrieval, ensuring that models could infer the subject category before searching for the answer. Additionally, annotators identified the ground truth document and the specific page from which the answer should be retrieved. The dataset was annotated in two rounds to ensure consistency and support models in evidence-based reasoning and retrieval across open-domain handwritten notes.

3.3.2. Dataset Statistics

NoTeS-Bank encompasses a comprehensive collection of 2,000 rigorously annotated QAs that are systematically apportioned, with 1,000 designated to **Task 1**, and an equivalent 1,000 ascribed to **Task 2**.

Table 2 visually presents the comprehensive statistics of the dataset. For the **Task 1**, the 649 distinct documents get used, which surges to `2,732` when incorporating distractor documents for the **Task 2**. Upon meticulous examination, the dataset reveals a repertoire of `1,982` unique questions juxtaposed with `1,583` distinct answers. A notable attribute of this data set is its composition of the diversified answer format: `1,244` responses are articulated in plain text, which makes up approximately `61.35%` of the total, while `758` responses are rendered in LaTeX or formulaic notation, representing the remaining `38.65%`. This format combines textual replies with scientific notations from diverse scholarly sources. The questions exhibit a maximum length of `48 tokens`, with an average of `12.4`. The answers are notably more concise, attaining a maximum length of `32` and a brief average of `3.9 tokens`, suggesting that numerous answers are direct or predominantly numerical.

4. Baseline Models

To evaluate performance on the Evidence-Based VQA task, we establish diverse baselines covering vision-language models (VLMs), OCR-based pipelines, and retrieval-augmented approaches. These baselines assess how different model architectures handle handwritten documents,

Table 3. Performance comparison of Open and Closed VLM-Based models, OCR + LLM, Layout + OCR + LLM, and VLM-Based OCR models on the Evidence-Based VQA task in NOTES-BANK. RL: Region-Level Layout; WL: Word-Level Layout

Model	#Param	Context Window	ANLS* (%)	IoU Metrics			Category Accuracy (%)	
				Avg IoU	IoU@5	IoU@10	Local	Global
Open VLM-Based Models								
Qwen-2.5-VL [16]	7B	32K	28.21	0.0136	0.0727	0.0518	11.83	4.86
Intern-2.5-VL [17]	8B	16k	21.44	0.0097	0.0490	0.0308	6.47	7.91
LLaVA-OneVision [30]	8B	60k	23.34	-	-	-	-	-
Closed VLM-Based Models								
GPT-4o [47]	-	128k	17.88	0.0122	0.0360	0.0381	12.00	9.00
Gemini 2.5 Pro [61]	-	2M	24.49	0.0130	0.0160	0.0367	0.2000	-
OpenAI o1 [48]	-	100k	15.06	0.0107	0.0565	0.0335	15.60	11.20
GPT-4.5 [1]	-	128k	21.65	0.0186	0.0898	0.0579	13.40	7.19
OCR + Closed LLM Models								
Google OCR [22] + Deepseek-R1 [34]	7B	128k	12.46	-	-	-	1.04	0.2283
OCR + Open LLM Models								
Google OCR [22] + Qwen2.5 [52]	7B	-	12.06	0	0	0	7.60	4.80
Amazon Textract [22] + Qwen2.5 [52]	7B	-	9.32	0	0	0	9.07	0.8240
Google OCR [22] + LLaMA 3.1 [15]	8B	-	9.40	0.0077	0.0486	0.0200	5.78	3.21
Amazon Textract [22] + LLaMA 3.1 [15]	8B	-	7.23	0.0032	0.0189	0.0101	-	-
Amazon Textract [22] + LLaMA 3.1 [15] + RL	8B	-	5.37	0	0	0	0.4124	0.6185
Amazon Textract [22] + LLaMA 3.1 [15] + WL	8B	-	7.54	0.0042	0.0261	0.0132	0.2062	1.64
Amazon Textract [22] + LLaMA 3.1 [15] + RL + WL	8B	-	7.01	0.0032	0.0065	0.0065	0.2061	1.44
VLM-Based OCR Models								
Nougat OCR [8] + LLaMA 3.1 [15]	-	-	3.20	-	-	-	0	0
GOT 2.0 OCR [69] + LLaMA 3.1 [15]	-	-	1.73	-	-	-	0	0
olmOCR [51] + LLaMA 3.1 [15]	-	-	4.86	0.0001	0.0011	0	0	0
Nougat OCR [8] + Qwen2.5 [52]	-	-	2.76	-	-	-	3.80	1.20
GOT 2.0 OCR [69] + Qwen2.5 [52]	-	-	9.59	-	-	-	3.40	0.8000
olmOCR [51] + Qwen2.5 [52]	-	-	4.69	-	-	-	2.60	1.60
Human Baseline	-	-	61.11	0.4009	0.5000	0.4312	83.16	79.34

reasoning, and evidence selection.

Vision-Language Models. VLMs holistically process visual and textual features, making them well-suited for handwritten notes where OCR struggles. We benchmark both *open* (Qwen-2.5-VL [16], Intern-2.5-VL [17], LLaVA [30]) and *closed* (GPT-4o [47], Gemini 2.5-Pro [61], GPT-4.5 [1], O1 [48]) models to analyze their multimodal reasoning capabilities.

OCR + LLM-Based Models. Traditional OCR pipelines extract text before reasoning, but handwriting often causes recognition errors. We evaluate Google OCR, Amazon Textract, and OLM OCR combined with LLaMA 3.1 [15] and Qwen2.5 [52] to measure OCR impact on answer quality.

Layout + OCR + LLM Models. Standard OCR pipelines discard document structure. We introduce layout-aware approaches incorporating region- and word-level cues (e.g., Textract + Layout Prompt) to assess document retrieval and reasoning improvements.

VLM-Based OCR Models. Instead of explicit text extraction, models such as Nougat OCR and GOT 2.0 OCR attempt direct transcription using vision-language understanding. These baselines highlight the limitations of OCR-free approaches in handwriting recognition. Models are compared across *ANLS** (answer similarity), *IoU* (evidence selection), and *category accuracy* (domain classification) to capture end-to-end document understanding. By selecting these baselines, we establish a comprehensive benchmark to drive advancements in handwritten document QA.

Human performance. For human evaluation, we collected responses from domain-aware individuals uninvolved in annotation. Each participant independently answered questions and marked evidence regions in the handwritten documents, ensuring unbiased assessment. Human responses consistently outperformed models in both accuracy and grounding—especially on complex, multimodal queries—underscoring the challenge of NOTES-BANK and the gap between model and expert-level understanding.

5. Experiments and Analysis

5.1. Evaluation Protocols

Prior work [73] has explored the reasoning capabilities of foundation models in visual tasks, primarily through qualitative analysis. In contrast, our objective with NOTES-BANK is to establish a systematic and unified evaluation protocol that enables both quantitative and qualitative assessment of vision-language models for symbolic and multimodal reasoning in handwritten scientific documents. We introduce a comprehensive benchmarking strategy for NOTES-BANK, encompassing both Evidence-Based VQA and Open-Domain QA). The models included in our benchmark range from OCR-enhanced LLMs to open and closed-source vision-language models, as detailed in Table 4. We report results across multiple evaluation dimensions, including answer correctness (ANLS*), evidence localization

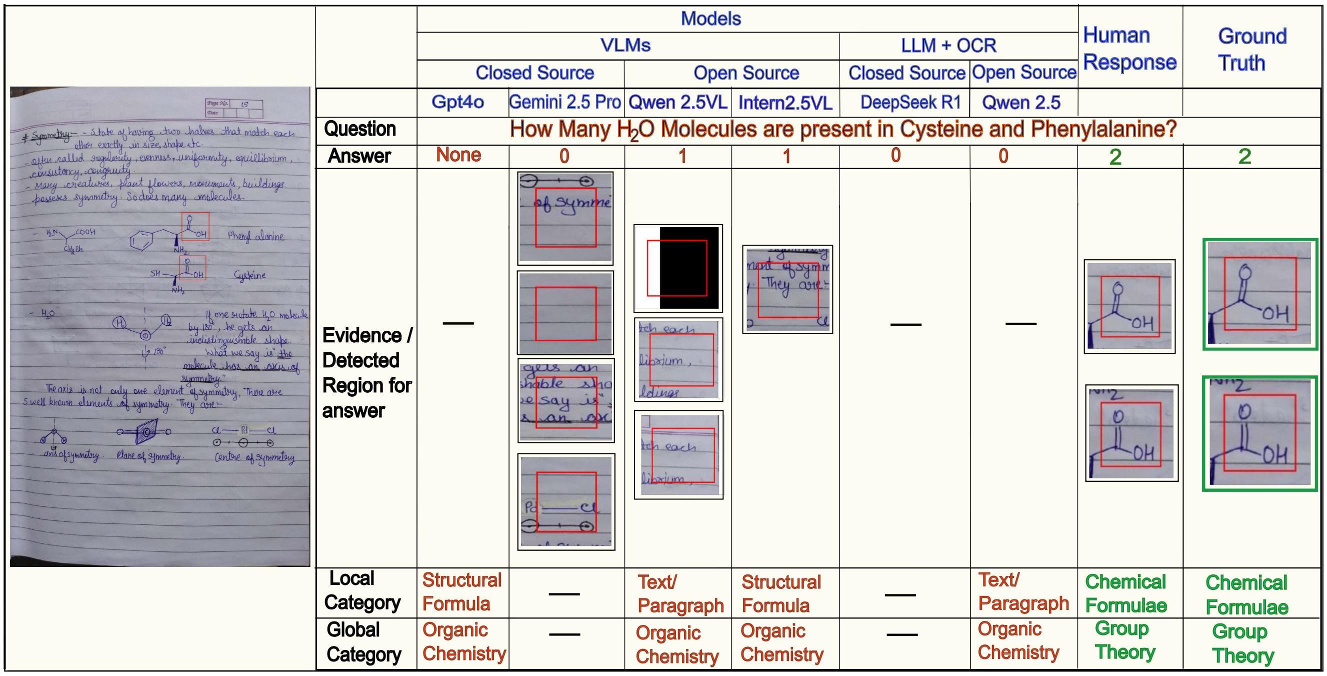

Figure 4. *Qualitative comparison of VLMs, OCR+LLMs, and human responses on NoTeS-Bank Evidence-Based VQA.* Highlights challenges in grounded answer retrieval from handwritten scientific notes, model limits in region detection and domain-specific reasoning also illustrates local (e.g., Structural Formula/Flowchart) and global (e.g., Organic Chemistry/Reproduction in Organisms) categories.

(IoU), document retrieval (Recall@K, MRR), and category prediction accuracy in Sec. 5.2.

In addition to quantitative performance metrics, we provide qualitative analyses in Figure 4 and Figure 5 of representative failure cases and model outputs, shedding light on limitations in layout reasoning, symbol understanding, and evidence attribution. Given the relatively stronger performance of GPT-4o [47] in multimodal tasks, we present comparisons against its peers, highlighting both its strengths and persistent challenges. Through this evaluation framework, NoTeS-Bank aims to serve as a diagnostic benchmark for the next generation of vision-language models in handwritten document understanding and retrieval.

5.2. Results and Discussion

Evaluations on *Evidence-Based* and *Open-Domain QA* in NoTeS-Bank reveal persistent challenges in handwritten document understanding. VLMs and multimodal LLMs, while promising, still falter on fine-grained reasoning, symbol interpretation, and multimodal retrieval.

Evidence-based VQA: For this task, models relying solely on OCR pipelines exhibit lower IoU and ANLS scores (shown in Table 3), reinforcing the limitation of text-only processing for handwritten notes. VLMs show improved performance by incorporating visual and structural cues, but evidence localization remains a major challenge.

Open-Domain QA: From Table 4, it is evident that multimodal retrieval-augmented generation (mmRAG) models outperform traditional text-based retrievers like BM25 [56] and DPR [27], particularly in Recall@5 and MRR metrics. However, even top models like ColPali and ColQwen [18] struggle with long-context retrieval over handwritten documents, highlighting the need for improved indexing over sparse visual information.

Table 4. Performance comparison of various methods. ANLS measures answer accuracy; R@1, MRR, R@5 and ACC (Global category accuracy) measure page retrieval.

Method	Accuracy	Page Retrieval			Domain
	ANLS* (%)	R@1	MRR	R@5	ACC (%)
Text-based RAG					
TF-IDF [56] + LLaMa 3.1 8B [15]	3.95	0.0180	0.0314	0.0580	5.80
BM 2.5 [56] + LLaMa 3.1 8B [15]	7.21	0.0340	0.0515	0.0820	7.60
Mp Net [58] + LLaMa 3.1 8B [15]	4.82	0.0200	0.0387	0.0760	6.60
Minilm [55] + LLaMa 3.1 8B [15]	4.01	0.0140	0.0245	0.0380	8.00
ColQwen [19] + Qwen2-VL (7B) [67]	34.19	0.2180	0.2430	0.2880	30.60
ColPali [19] + Qwen2-VL (7B) [67]	32.94	0.2120	0.2430	0.2900	30.40
Human Baseline	86.67	0.8125	0.8125	–	28.99

Impact of OCR on Document Understanding: OCR-based methods perform significantly worse in both tasks, particularly for handwritten mathematical equations, symbols, and complex scientific notations. Models such as Google OCR and Textract OCR + LLaMA 3.1 suffer from: ***(i)*** Loss of spatial and semantic relationships is crucial for layout-heavy content. ***(ii)*** Difficulty in transcribing non-standard handwritten characters. ***(iii)*** Inability to provide reliable evidence grounding due to segmentation errors.

This highlights the limitations of treating handwritten document QA as a text-only problem, reinforcing the ne-

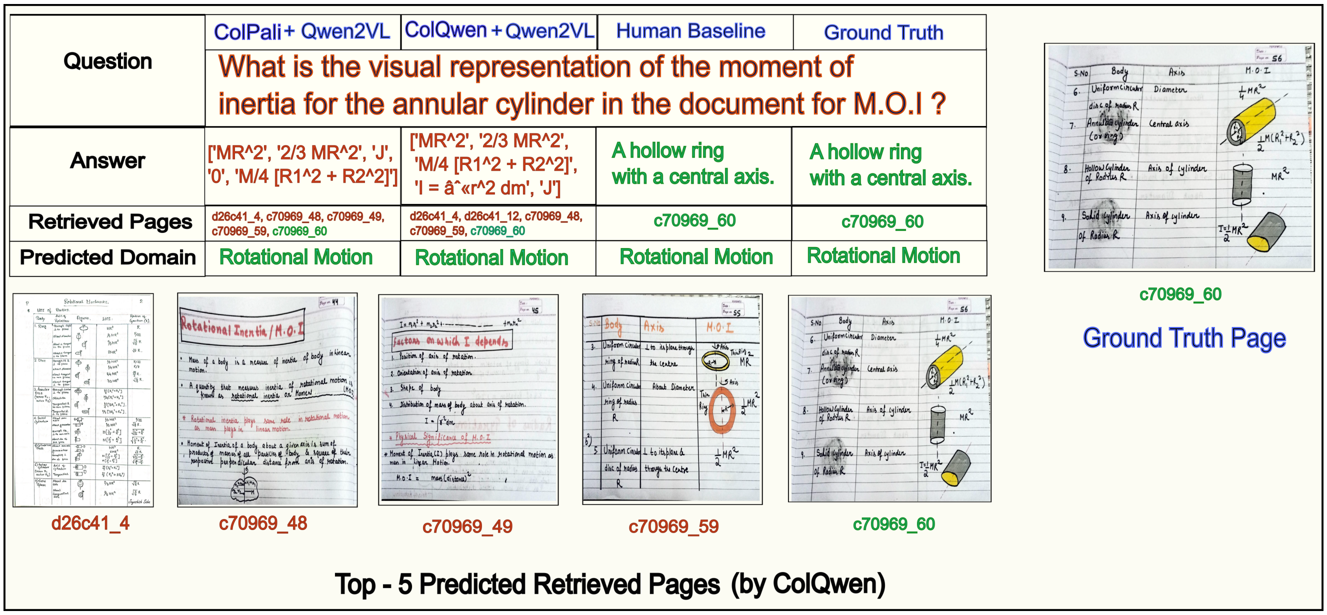

Figure 5. *Qualitative comparison of Open-Domain VQA performance in NoTeS-Bank benchmark.* Each query requires retrieving relevant handwritten pages from a large corpus and reasoning across them. The figure depicts VLM predictions (ColPali + Qwen2VL, ColQwen + Qwen2VL), human responses, and ground-truth. Differences in answers, documents, and domains highlight retrieval, inference, and multimodal reasoning challenges over noisy visuals.

cessity for joint vision-language reasoning.

Vision-Language Models and the Multimodal Challenge: While closed VLMs (GPT-4o [47], Gemini 2.5-Pro [61]) outperform open VLMs in answer generation, both struggle with localizing relevant evidence. Intern-2.5-VL [17] and Qwen-2.5-VL [16] show promise in handling handwritten content but fail to generalize across different domain categories. For multimodal retrieval, ColPali and ColQwen [19] improve retrieval accuracy but still fail to effectively fuse retrieved document context into reasoning steps. This suggests the need for better cross-modal pretraining strategies that explicitly model symbolic and spatial dependencies.

5.3. Error Analysis

Hallucinations on Unanswerable Questions: Unanswerable questions assess the model's proper comprehension of the document and its propensity for generating hallucinations. Illustrative examples are outlined in Figure 10 in the supplementary. In such cases, the expected response should constitute a phrase along the lines "cannot be answered from the given context". However, models underperform compared to humans, often generating fabricated responses, inaccurate evidence via erroneous bounding boxes, and wrong answers.

Weak Layout Awareness: The Models exhibit significant challenges in accurately interpreting handwritten equations in Mathematics, Chemistry, Diagrams, and Tabular data. In the 1st task, 61% of errors arise from non-textual components (Figure 2 in the supplementary). Key failing subjects: Operating Systems, Reproductive Processes in Organisms, Design and Analysis of Algorithms, and Group Theory in Chemistry (Figures 1, 3 in the supplementary). The 2nd task exhibits similar failures in these areas, showing consistent difficulties (Figure 4 in the supplementary).

Failure to Retrieve Key Evidence: Even the most proficient models often erroneously retrieve incorrect documents, culminating in incomplete responses. Empirical observations indicate that the state-of-the-art retrieval system, such as ColQwen, frequently selects an inaccurate document as the first entry among the top five retrieved documents. This phenomenon is particularly pronounced when the model engages with specialized disciplines, including Electrical Materials, Mechanical Properties of Solids, and Fluid Mechanics (Figures 11 and 12 of supplementary).

6. Conclusion

We presented NOTES-BANK, a novel benchmark designed to evaluate multimodal reasoning over unstructured handwritten scientific notes, addressing a critical gap in current visual document understanding tasks. Through rigorous evaluations across state-of-the-art VLM/LLMs and retrieval methods, we revealed significant limitations in handling multimodal content, visual grounding, and complex retrieval scenarios, particularly for unstructured, informal visio-graphical documents, where answers depend on interpreting diagrams, equations, or layout cues absent from OCR transcriptions. Our analysis underscores the need for models that effectively fuse non-trivial visual cues with textual information beyond traditional OCR-based approaches. Thus, NoTeS-Bank establishes a challenging new benchmark to push multimodal document understanding research.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 3, 6

[2] Rajas Agashe, Srinivasan Iyer, and Luke Zettlemoyer. JuICe: A large scale distantly supervised dataset for open domain context-based code generation. 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)*, pages 5436–5446, Hong Kong, China, 2019. Association for Computational Linguistics. 3

[3] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 3

[4] Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R Manmatha. Docformer: End-to-end transformer for document understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 993–1003, 2021. 2

[5] Gilles Baechler, Srinivas Sunkara, Maria Wang, Fedir Zubach, Hassan Mansoor, Vincent Etter, Victor Cărbune, Jason Lin, Jindong Chen, and Abhanshu Sharma. Screenai: A vision-language model for ui and infographics understanding. *arXiv preprint arXiv:2402.04615*, 2024. 3

[6] Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, et al. Longbench: A bilingual, multitask benchmark for long context understanding. *arXiv preprint arXiv:2308.14508*, 2023. 4

[7] Johannes Bjerva, Nikita Bhutani, Behzad Golshan, Wang-Chiew Tan, and Isabelle Augenstein. SubjQA: A Dataset for Subjectivity and Review Comprehension. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5480–5494, Online, 2020. Association for Computational Linguistics. 3

[8] Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. Nougat: Neural optical understanding for academic documents. *arXiv preprint arXiv:2308.13418*, 2023. 2, 6

[9] Soumitri Chattopadhyay, Sanket Biswas, Emanuele Vivoli, and Josep Lladòs. Towards generative class prompt learning for fine-grained visual recognition. In *British Machine Vision Conference*, 2024. 3

[10] Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric Xing, and Liang Lin. GeoQA: A geometric question answering benchmark towards multimodal numerical reasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 513–523, Online, 2021. Association for Computational Linguistics. 3

[11] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 3

[12] Jaemin Cho, Debanjan Mahata, Ozan Irsoy, Yujie He, and Mohit Bansal. M3docrag: Multi-modal retrieval is what you need for multi-page multi-document understanding. *arXiv preprint arXiv:2411.04952*, 2024. 2, 3, 4

[13] Alloy Das, Sanket Biswas, Umapada Pal, and Josep Lladós. Diving into the depths of spotting text in multi-domain noisy scenes. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 410–417. IEEE, 2024. 3

[14] Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. Transvg: End-to-end visual grounding with transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1749–1759, 2021. 2

[15] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 2, 6, 7

[16] Shuai et. al. Qwen2.5-vl technical report, 2025. 6, 8

[17] Zhe et. al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling, 2025. 6, 8

[18] Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. Colpali: Efficient document retrieval with vision language models. In *The Thirteenth International Conference on Learning Representations*, 2024. 2, 7

[19] Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, CELINE HUDELOT, and Pierre Colombo. Colpali: Efficient document retrieval with vision language models. In *The Thirteenth International Conference on Learning Representations*, 2025. 7, 8

[20] Lin Gui, Jiannan Hu, Yulan He, Ruifeng Xu, Qin Lu, and Jiachen Du. A question answering approach for emotion cause extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1593–1602, Copenhagen, Denmark, 2017. Association for Computational Linguistics. 3

[21] Seungju Han, Kavel Rao, Allyson Ettinger, Liwei Jiang, Bill Yuchen Lin, Nathan Lambert, Yejin Choi, and Nouha Dziri. Wildguard: Open one-stop moderation tools for safety risks, jailbreaks, and refusals of LLMs. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. 5

[22] Thomas Hegghammer. Ocr with tesseract, amazon textract, and google document ai: a benchmarking experiment. *Journal of Computational Social Science*, 5(1):861–882, 2022. 2, 6

[23] Mark Hopkins, Ronan Le Bras, Cristian Petrescu-Prahova, Gabriel Stanovsky, Hannaneh Hajishirzi, and Rik Koncel-Kedziorski. SemEval-2019 task 10: Math question answering. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 893–899, Minneapolis, Minnesota, USA, 2019. Association for Computational Linguistics. 3

[24] Xuming Hu, Zhijiang Guo, GuanYu Wu, Aiwei Liu, Lijie Wen, and Philip Yu. CHEF: A pilot Chinese dataset for evidence-based fact-checking. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3362–3376, Seattle, United States, 2022. Association for Computational Linguistics. 3

[25] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. Layoutlmv3: Pre-training for document ai with unified text and image masking. *arXiv preprint arXiv:2204.08387*, 2022. 2, 3

[26] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. PubMedQA: A dataset for biomedical research question answering. 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)*, pages 2567–2577, Hong Kong, China, 2019. Association for Computational Linguistics. 3

[27] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online, 2020. Association for Computational Linguistics. 7

[28] Yoonsik Kim, Moonbin Yim, and Ka Yeon Song. Tablevqa-bench: A visual question answering benchmark on multiple table domains. *arXiv preprint arXiv:2404.19205*, 2024. 3

[29] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *Advances in Neural Information Processing Systems*, 37:87874–87907, 2024. 2

[30] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 2, 6

[31] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 3

[32] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 3

[33] Minghao Li, Yiheng Xu, Lei Cui, Shaohan Huang, Furu Wei, Zhoujun Li, and Ming Zhou. Docbank: A benchmark dataset for document layout analysis, 2020. 1

[34] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024. 2, 6

[35] Chenxiao Liu and Xiaojun Wan. CodeQA: A question answering dataset for source code comprehension. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2618–2632, Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. 3

[36] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 3

[37] Jiahua Liu, Yankai Lin, Zhiyuan Liu, and Maosong Sun. XQA: A cross-lingual open-domain question answering dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2358–2368, Florence, Italy, 2019. Association for Computational Linguistics. 3

[38] Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3622–3628. International Joint Conferences on Artificial Intelligence Organization, 2020. Main track. 3

[39] Shayne Longpre, Yi Lu, and Joachim Daiber. Mkqa: A linguistically diverse benchmark for multilingual open domain question answering, 2020. 3

[40] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023. 2, 3

[41] Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, et al. Mmlongbench-doc: Benchmarking long-context document understanding with visualizations. *arXiv preprint arXiv:2407.01523*, 2024. 2, 3, 4

[42] Minesh Mathew, Ruben Tito, Dimosthenis Karatzas, R Manmatha, and CV Jawahar. Document visual question answering challenge 2020. *arXiv preprint arXiv:2008.08899*, 2020. 1, 3

[43] Minesh Mathew, Lluis Gomez, Dimosthenis Karatzas, and CV Jawahar. Asking questions on handwritten document collections. *International Journal on Document Analysis and Recognition (IJDAR)*, 24(3):235–249, 2021. 1

[44] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706, 2022. 3

[45] Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. Ambigqa: Answering ambiguous open-domain questions, 2020. 3

[46] Anastasios Nentidis, Georgios Katsimpras, Eirini Vandorou, Anastasia Krithara, Antonio Miranda-Escalada, Luis Gasco, Martin Krallinger, and Georgios Paliouras. Overview of BioASQ 2022: The tenth BioASQ challenge on large-

scale biomedical semantic indexing and question answering. In *Lecture Notes in Computer Science*, pages 337–361. Springer International Publishing, 2022. 3

[47] OpenAI and Aaron et. al. Gpt-4o system card, 2024. 2, 5, 6, 7, 8

[48] OpenAI, :, and Aaron Jaech et al. Openai o1 system card, 2024. 6

[49] David Peer, Philemon Schöpf, Volckmar Nebendahl, Alexander Rietzler, and Sebastian Stabinger. Anls* – a universal document processing metric for generative large language models, 2025. 4

[50] Birgit Pfitzmann, Christoph Auer, Michele Dolfi, Ahmed S Nassar, and Peter Staar. Doclaynet: A large human-annotated dataset for document-layout segmentation. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pages 3743–3751, 2022. 2

[51] Jake Poznanski, Jon Borchardt, Jason Dunkelberger, Regan Huff, Daniel Lin, Aman Rangapur, Christopher Wilhelm, Kyle Lo, and Luca Soldaini. olmocr: Unlocking trillions of tokens in pdfs with vision language models. *arXiv preprint arXiv:2502.18443*, 2025. 2, 6

[52] Qwen, :, An Yang, Baosong Yang, and Beichen et. al. Qwen2.5 technical report, 2025. 6

[53] Preethi Raghavan, Jennifer J Liang, Diwakar Mahajan, Rachita Chandra, and Peter Szolovits. emrKBQA: A clinical knowledge-base question answering dataset. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 64–73, Online, 2021. Association for Computational Linguistics. 3

[54] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016. 2

[55] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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)*, pages 3982–3992, Hong Kong, China, 2019. Association for Computational Linguistics. 7

[56] Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4), 2009. 7

[57] Juan A Rodriguez, Xiangru Jian, Siba Smarak Panigrahi, Tianyu Zhang, Aarash Feizi, Abhay Puri, Akshay Kalkunte Suresh, François Savard, Ahmed Masry, Shravan Nayak, et al. Bigdocs: An open dataset for training multimodal models on document and code tasks. In *The Thirteenth International Conference on Learning Representations*, 2025. 1

[58] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: masked and permuted pre-training for language understanding. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2020. Curran Associates Inc. 7

[59] Manan Suri, Puneet Mathur, Franck Dernoncourt, Kanika Goswami, Ryan A Rossi, and Dinesh Manocha. Visdom: Multi-document qa with visually rich elements using multimodal retrieval-augmented generation. *arXiv preprint arXiv:2412.10704*, 2024. 2, 3, 4

[60] Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. Slidevqa: A dataset for document visual question answering on multiple images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13636–13645, 2023. 3

[61] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 2, 6, 8

[62] Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. Document collection visual question answering. In *Document Analysis and Recognition–ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part II 16*, pages 778–792. Springer, 2021. 1, 3

[63] Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. Hierarchical multimodal transformers for multi-page docvqa. *arXiv preprint arXiv:2212.05935*, 2022. 4

[64] Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. Newsqa: A machine comprehension dataset. *arXiv preprint arXiv:1611.09830*, 2016. 2

[65] Jordy Van Landeghem, Lukasz Borchmann, Rubèn Tito, Michał Pietruszka, Dawid Jurkiewicz, Rafał Powalski, Paweł Józiak, Sanket Biswas, Mickaël Coustaty, and Tomasz Stanisławek. ICDAR 2023 Competition on Document UnderstanDing of Everything (DUDE). In *Proceedings of ICDAR 2023*, 2023. 3

[66] Jordy Van Landeghem, Rubèn Tito, Łukasz Borchmann, Michał Pietruszka, Pawel Joziak, Rafal Powalski, Dawid Jurkiewicz, Mickaël Coustaty, Bertrand Anckaert, Ernest Valveny, et al. Document understanding dataset and evaluation (dude). In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19528–19540, 2023. 1, 3, 4

[67] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 2, 7

[68] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. In *Advances in Neural Information Processing Systems*, pages 95266–95290, 2024. 5

[69] Haoran Wei, Chenglong Liu, Jinyue Chen, Jia Wang, Lingyu Kong, Yanming Xu, Zheng Ge, Liang Zhao, Jianjian Sun, Yuang Peng, et al. General ocr theory: Towards ocr-2.0 via a unified end-to-end model. *arXiv preprint arXiv:2409.01704*, 2024. 2, 6

[70] Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. Layoutlm: Pre-training of text and layout

for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1192–1200, 2020. 2, 3

[71] Li Yang, Yan Xu, Chunfeng Yuan, Wei Liu, Bing Li, and Weiming Hu. Improving visual grounding with visual-linguistic verification and iterative reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2

[72] Yi Yang, Wen-tau Yih, and Christopher Meek. WikiQA: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2013–2018, Lisbon, Portugal, 2015. Association for Computational Linguistics. 3

[73] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1):1, 2023. 6

[74] Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, and Maosong Sun. VisRAG: Vision-based retrieval-augmented generation on multi-modality documents. In *The Thirteenth International Conference on Learning Representations*, 2025. 2

[75] Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. Reclor: A reading comprehension dataset requiring logical reasoning. In *International Conference on Learning Representations (ICLR)*, 2020. 3

[76] Majid Zarharan, Mahsa Ghaderan, Amin Pourdabiri, Zahra Sayedi, Behrouz Minaei-Bidgoli, Sauleh Eetemadi, and Mohammad Taher Pilehvar. ParsFEVER: a dataset for Farsi fact extraction and verification. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 99–104, Online, 2021. Association for Computational Linguistics. 3

[77] Qiyuan Zhang, Lei Wang, Sicheng Yu, Shuohang Wang, Yang Wang, Jing Jiang, and Ee-Peng Lim. NOAHQA: Numerical reasoning with interpretable graph question answering dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4147–4161, Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. 3

[78] Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Khai Hao, Xu Han, Zhen Leng Thai, Shuo Wang, Zhiyuan Liu, et al. Infinitybench: Extending long context evaluation beyond 100k tokens. *arXiv preprint arXiv:2402.13718*, 2024. 4

[79] Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. Publaynet: largest dataset ever for document layout analysis. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1015–1022. IEEE, 2019. 1

[80] Fengbin Zhu, Ziyang Liu, Xiang Yao Ng, Haohui Wu, Wenjie Wang, Fuli Feng, Chao Wang, Huanbo Luan, and Tat Seng Chua. Mmdocbench: Benchmarking large vision-language models for fine-grained visual document understanding. *arXiv preprint arXiv:2410.21311*, 2024. 1, 3